UNIVERSITÉ PIERRE ET MARIE CURIE

THÈSE

pour obtenir le grade de

DOCTEUR de Université Pierre et Marie Curie

Spécialité : **Sciences de l'Environnement**

préparée au laboratoire **CNRS/LATMOS - LSCE/IPSL**

dans le cadre de l'École Doctorale **Des Sciences de l'Environnement / ED129**

présentée et soutenue publiquement

par

M. Gazeaux Julien

le 07 février 2011

Titre:

Méthodes probabilistes d'extraction de signaux cachés appliquées à des problémes de sciences de l'atmosphére

Directeur de thèse: **Bekki Slimane**

Co-directeur de thèse: **Naveau Philippe**

Jury

Dr. Allard Denis,	Rapporteur
Dr. Mestre Olivier,	Rapporteur
Dr. Cosme Emmanuel,	Examinateur
Dr. De Mazière Martine,	Examinatrice
Pr. Ravetta Francois,	Examinateur
Pr. Thiria Sylvie,	Examinatrice
Dr. Bekki Slimane,	Directeur de thèse
Dr. Naveau Philippe,	Co-directeur de thèse

Résumé

Ce travail de thèse est consacré à la problématique de l'extraction de signaux dans le domaine des sciences de l'atmosphère. Le point commun des problèmes considérés est la notion de détection et d'estimation de signaux cachés. L'approche par la modélisation probabiliste s'est avérée y être bien adaptée.

Nous nous sommes attachés à répondre à différentes questions telles que : quel type d'information s'attend-on à trouver dans un jeu de données ? Le signal supposé caché se trouve-t-il réellement dans les données d'étude ? Comment détecter l'instant d'occurrence d'un phénomène, comment le caractériser (timing, amplitude ...) ? Si un tel signal est détecté, quelle (in)certitude est associée à cette détection ? Nous répondons à ces différentes questions au travers du développement de différents modèles probabilistes de détection d'évènements cachés. Nous nous sommes intéressés à différents modèles non stationnaires, dont deux sont notamment présentés dans un cadre multivarié.

Au travers de trois modèles probabilistes décrivant des signaux cachés divers (rupture de variance, signaux éruptifs et changement de moyenne), nous avons aussi développé des méthodes associées de détection. Le premier modèle est appliqué à la détection de nuages stratosphériques polaires dans des profils lidar, le deuxième à des éruptions volcaniques dans des séries chronologiques de sulfate et enfin le troisième est appliqué à la détection de la date de l'onset de la mousson Africaine dans des données géophysiques liées à la dynamique atmosphérique et aux précipitations. Les différentes méthodes mises en place font appel à une variété de techniques de modélisation probabiliste allant de la maximisation de rapport de vraisemblance associée à des tests d'hypothèses à la résolution de filtres de Kalman dans un cadre non stationnaire et non linéaire pour la décomposition de séries multivariées couplée à la détection des signaux cachés. Les difficultés techniques liées à l'extraction de signaux cachés sont analysées et les performances des différents algorithmes sont évaluées. Les résultats obtenus confirment l'intérêt des méthodes probabilistes appliquées à ces problématiques de signaux cachés en sciences de l'atmosphère.

Abstract

This work is devoted to the study of signal extraction applied to atmospheric sciences. The main thread in the different studies presented here is the detection, estimation and characterization of hidden signals. The probabilistic modeling approach has turned out to be well suited to this problematic.

For all the problems considered here, the main objective was to respond to the following questions : Which type of information do we expect to find in our data sets ? Are expected hidden signals actually present in the data sets ? How is it possible to detect the time of occurrence of a phenomenon ? How can it be characterized (timing, amplitude ...) ? If such a signal is detected, what is the uncertainty associated to the detection ? These questions are tackled through three probabilistic hidden signals models. We have focused on non stationary and multivariate models.

Along three probabilistic models describing diverse hidden signals (break of variance, pulse-like signals and shift in the mean), we have also developed associated detection methods. The first model is applied to the detection of polar stratospheric clouds in lidar profiles, the second to volcanic eruptions in time series of sulfate and finally the third is applied to detect the date of the onset of the African monsoon in data related to atmospheric dynamics and precipitation. The various methods use a range of techniques in probabilistic modeling, from the likelihood ratio maximization associated with hypothesis testing to the resolution of Kalman filters in a non stationary and non linear framework for the decomposition of multivariate series coupled with the detection of hidden signals. The technical difficulties associated to the extraction of hidden signals are analyzed and the performances of the various algorithms are estimated. The results confirm the interest and the potential of probabilistic methods applied to problems of hidden signals in atmospheric sciences.

Remerciements

Exercice facile et délicat, pour lequel il n'y a pas besoin de chercher longtemps pour trouver, mais dont une réponse exhaustive, but ultime est pourtant impossible.

Je remercie pour leurs soutiens, leurs patiences et leurs confiances lors de ces trois années mes directeurs de thèse, Slimane et Philippe. Un travail de thèse tient souvent à un bon équilibre entre l'experience et les attentes de l'encadrant d'une part et les choix, plus ou moins avisés, plus ou moins inspirés du thésitif d'autre part.

Slimane, je dois beaucoup à tes connaissances et ton recul sur les questions liées à l'atmosphère, pour ton bouillonnement d'idées et tes encouragements. Philippe, je te remercie de m'avoir guider dans les problèmes statistiques et permis d'avancer aussi rapidement. Je vous suis particulièrement reconnaissant de m'avoir dirigé, conseillé tout en me laissant la responsabilité de mon travail, dans une atmosphère de confiance idéale à l'achèvement de ce manuscrit. J'espère sincèrement, à l'avenir, pouvoir continuer à collaborer avec vous.

Je souhaite tout autant remercier ma famille, mon père, ma mère, mes soeurs qui ont su me rassurer, quand il le fallait et m'amener à me convaincre, par leurs questions, du sens de mes choix.

Je tiens à remercier les membres du jury, tout spécialemnent Denis Allard et Olivier Mestre, qui par le temps consacré à la relecture et leurs rapports ont notablement contribué à ce manuscrit. Je remercie Emmanuel Cosme, Martine De Mazière, Francois Ravetta et Sylvie Thiria d'avoir examiné la thèse et participé à la soutenance.

Je souhaite également remercier Sophie Godin-Beckman et Monique Petitdidier tout d'abord pour m'avoir embauché en tant qu'ingénieur quelques temps avant le début de ma thèse et m'avoir ainsi permis de découvrir le LATMOS et le monde de la recherche. De chaleureux remerciements à M. Jacques Portes, qui m'a enseigné la programmation en master et dont nos conversations ont grandement contribuées à me conforter dans mes choix.

J'ai eu la chance de travailler entre le LSCE/CEA et le LATMOS/CNRS, deux décors, deux ambiances complémentaires. D'un côté, Gyf-Sur-Yvette et ses grands espaces, sa

forêt, ses matchs de touch-rugby du midi et son RER capricieux. Merci à Julien et Jérôme pour m'avoir fait un peu de place à leurs côtés. De l'autre, Paris, Jussieu, sa tour Zamansky, que j'ai longtemps scrutée depuis ma chaise pour y chercher l'inspiration, et une équipe accueillante et motivante. Merci entre autre à Christophe, José, Fabien, Rémy et Thomas pour les sorties de bureau à l' "accadémie" qui m'ont permis, bien des fois, de m'échapper un peu... Je pense bien évidemment également à Anne, Ariela, et Juliette pour votre amitié, et la place que vous m'avez faite dans vos bureau et à bien d'autres encore. Une pensée particulière à Maya, qui à su me supporter patiemment pendant toutes ces années, malgré mes gesticulations tout au long de la journée. J'espère que les pains aux chocolats auront adoucis un peu mon séjour...

Enfin, une pensée sincère à mes compagnons de galère et amis, ceux avec qui j'ai partagé mon quotidien lors de ces trois années : Manher, Manouko, RoBo, Tho, Zouzou, un profond merci...

pour vous deux ...

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction Générale

Ce chapitre d'introduction présente un éclairage sur l'utilisation des statistiques et probabilités dans les sciences de l'atmosphère, les différents jeux de données considérés durant cette thèse, ainsi que leurs caractéristiques. Nous y expliquons comment la réflexion probabiliste apporte un complément à l'étude déterministe de l'atmosphère. L'approche probabiliste, travaillant sur l'aléa, permet de mettre en évidence des informations qui ne sont pas directement accessibles. Cette problématique des signaux cachés est au coeur de ce travail de thèse, nous l'introduisons dans ce chapitre, et reviendrons plus en détails dessus au chapitre suivant.

Plan du Chapitre 1

1. **L'étude probabiliste en sciences de l'atmosphère**
2. **Les données géophysiques**
3. **L'apport de l'analyse probabiliste**
4. **L'extraction de données**
5. **Plan du manuscrit**

1 L'étude probabiliste en sciences de l'atmosphère

à partir de la fin du XIXème siècle s'est développé l'intérét de la vision probabiliste d'un monde que les lois de Newton peinaient à expliquer. L'hypothèse soutenue par la physique statistique de l'existence des molécules introduit l'idée que le monde *microscopique* (en considérant un grand nombre de particules) permettrait d'expliquer les systèmes *macroscopiques*. L'exemple historique est l'étude de l'équation de la chaleur qui est définie par l'interaction de nombreuses particules aux comportements aléatoire, ainsi [Perrin, 1913] explique que la chaleur ne peut être évaluée au niveau microscopique uniquement à partir de considérations déterministes. L'étude des lois de probabilités ouvre la voie à la compréhension de systémes composés d'une multitude d'éléments interagissant de maniére aléatoire. C'est sous l'impulsion de grands noms de la physique que cette idée s'est développée. Rudolf Clausius dépoussiérant les travaux de Sadi Carnot, James Maxwell décrivant une méthode statistique pour expliquer la cinétique des gaz (e.g. [Maxwell, 1866]), ou encore Ludwig Boltzmann, la méme année qui tenta d'imposer sa vision atomistique à des confréres hostiles ont déroulé les prémices de la physique statistique, qui seront formalisées au début des années 1900 par les travaux de Max Planck et Albert Einstein, notamment sur l'étude de la répartition statistique des particules (e.g. [Einstein, 1905]). De nombreux ouvrages (e.g. [Taton, 1995], [Mankiewicz, 2001]) relatent ces découvertes qui ont alimenté, jusqu'é nos jours la recherche en sciences de l'atmosphère, concept générique ayant pour objectif la compréhension de la réalité physique et chimique de l'atmosphère, de ses interactions avec la Terre, et l'activité humaine et englobant de nombreuses compétences scientifiques diverses, telles que la physique, la chimie, la biologie, la géologie, les mathématiques...

Plus récemment, les essors conjoints de l'intérét pour les sciences de l'atmosphère, des moyens de calculs et de stockage et le développement des instruments de mesure ont amené la recherche dans ce domaine à manipuler des systémes complexes de dimension croissante. La multiplication des différents modèles numériques climatiques (modèles régionaux, modèles globaux ...), et des systémes d'observation de l'atmosphère (satellites, lidar, ballons-sondes ...) ont parallélement conduit à la production de grands ensembles de données sur des phénoménes physiques complexes dont l'étude nécessite le développement de nouvelles méthodes d'analyse.
L'approche, au milieu du XXéme siécle, des sciences de l'atmosphère par des méthodes statistiques et probabilistes a ouvert de nouvelles applications aux mathématiques. Les répercussions des travaux du statisticien et physicien Sir Gilbert Walker en sont une

bonne illustration, [Katz, 2002] revient sur ce chercheur qui mit son nom à la fois sur les équations de Yule-Walker (qui résolvent l'estimation des paramètres d'un processus autorégressif, processus défini plus loin dans cette thèse) et sur la circulation de Walker (qui donne une explication de l'oscillation australe, elle-méme reliée au phénoméne El Niño, voir [Delmas et al., 2005]). Certains événements de la physique de l'atmosphère ne seront désormais plus considérés que d'un point de vue purement dynamique. Des ouvrages consacrés exclusivement à ce sujet apparaissent, nous citerons par exemple [Von Storch and Zwiers, 1999]. Un événement climatique observé, à un instant donné, est alors considéré comme une réalisation particuliére d'un ensemble d'événements climatiques possibles. Les phénoménes physiques ne sont plus étudiés de maniére statique, mais par le truchement des événements qui auraient également pu se produire à leurs places.

L'étude de séries, encore appelées séries chronologiques lorsque celles-ci sont indexées par le temps, englobe la description, l'analyse et la prédiction de séries de données ordonnées (e.g. [Brockwell and Davis, 2002]). Nous nous sommes attachés, dans ce travail de thèse, à la description et l'analyse de séries indexées soit par le temps (Chapitres 4 et 5) soit par l'altitude (Chapitre 3) et n'avons pas abordé les problématiques liées à la prédiction. Les modèles probabilistes et les méthodes d'inférence étudiés dans les différents développement de ce travail se concentrent essentiellement sur la recherche et l' extraction de signaux cachés, cette notion sera explicitée dans la suite du texte.

2 Les données géophysiques

Il existe de multiples sources de données en géophysique. Une distinction commune est de séparer les données issues des modèles numériques, tels que les GCM (Global Climate Model), des données d'observation. Les mesures atmosphériques proviennent d'instruments trés variés, nous citerons par exemple, les mesures par ballons auxquels sont attachés des capteurs radiosondes, les mesures par télédétections utilisant la déflexion d'ondes électromagnétiques (mesure Radar, Lidar), les instruments de mesure pouvant aussi être embarqués sur des plateformes, des bouées, par bateaux ou par avions (voir les Figures 1.1 é 1.4). Pour plus d'informations et détails sur les différents types de données existants nous conseillons la lecture du livre de [Delmas et al., 2005].

Dans ce travail de thèse, nous avons été amenés à étudier exclusivement des données d'observation. Il est pourtant intéressant d'étudier les différences entre les deux types de données pour en extraire la spécificité des données traitées. Plusieurs différences sont à

noter. Tout d'abord, les données des modèles numériques, cohérentes avec la physique du modèle, sont reproductibles, elles sont généralement disponibles en plus grande quantité, et se présentent sous forme de grilles spatio-temporelles réguliéres souvent ajustables par le modélisateur, en fonction des besoins. Les données d'observation, souvent échantillonnées de maniére irréguliére, sont des données qui ne sont pas parfaitement reproductibles (une observation se fait à un instant donné dans des conditions uniques). Elles présentent une variabilité généralement plus élevée que les données des modèles numériques. En effet, à la variabilité propre observée des données s'ajoutent les erreurs de mesure, de plus les modèles numériques par définition non exacts ne permettent pas de reproduire la totalité de la variabilité naturelle d'un signal, et sont donc généralement plus lisses. Ces deux raisons font que les données d'observation sont généralement plus bruitées que les données des modèles numériques. Ce phénoméne de bruit dans les données peut être aisément mis en évidence par un examen visuel de la densité spectrale de puissances des données (le module de la transformée de Fourier divisé par le temps d'intégration). On pourra ainsi voir ([Mahowald et al., 2003] par exemple) que les données d'observation présentent généralement une variabilité plus forte.

Néanmoins ces jeux de données sont complémentaires, chacun apportant des informations propres. Si les données d'observation témoignent de la réalisation d'un phénoméne physique, les données de modèle témoignent quant à elles de la connaissance de ce phénoméne. Des méthodes statistiques, telles que les méthodes d'assimilation permettent de combiner de maniére optimale les différents jeux de données afin d'en exploiter leurs intéréts spécifiques. De nombreux travaux d'Olivier Talagrand sont axés sur ces problématiques, on citera de maniére non exhaustive : [Talagrand, 2003] et tout récemment [Pires et al., 2010]. Parmi les différentes recherches auxquelles se sont intéressés ces auteurs, le filtre de Kalman tient une place importante. Il est également une pierre angulaire de ce manuscrit, et sera développé dans les Chapitres 4 et 5.

Dans ce travail de thèse, nous avons utilisé différentes sources de données d'observation. Pour l'étude des nuages stratosphériques polaires (PSC pour Polar Stratospheric Clouds) du Chapitre 3, nous avons utilisé des profils de rétrodiffusion lidar mesurés au dessus de la station scientifique de Dumont D'Urville en Antarctique ([Lachlan-Cope et al., 2009]). L'analyse des "pulses" volcaniques (Chapitre 4) a nécessité l'utilisation de données d'observations issues de forages de carottes de glace récupérées sur cinq sites géographiques différents du Groenland ([Gao et al., 2007]). Enfin, l'étude des ruptures dans la Mousson de l'Afrique de l'Ouest (Chapitre 5) s'est faite via des données de front

FIGURE 1.1 – Carotte de glace du Déme Fuji, en Antarctique, avec la téte du foret. Cette glace a été extraite d'une profondeur de 1332 métres. Elle a été déposée lé, il y a environ 89 000 ans. Illustration : © Dr. Hideaki Motoyama, Institut national de recherche polaire, Japon

FIGURE 1.2 – Ballon emportant une sonde à ozone. On peut distinguer de haut en bas : le ballon, le parachute (rouge), le réflecteur radar et enfin la nacelle. © SA-IPSL.

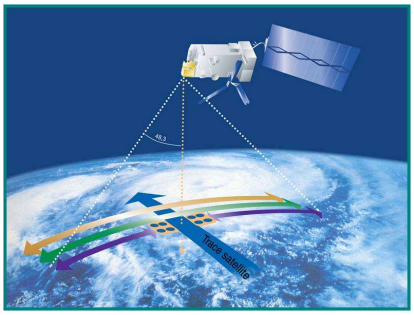

FIGURE 1.3 – Procédé d'acquisition par défilement de l'instrument IASI placé à bord du satellite MetOp depuis 2006. © CNES

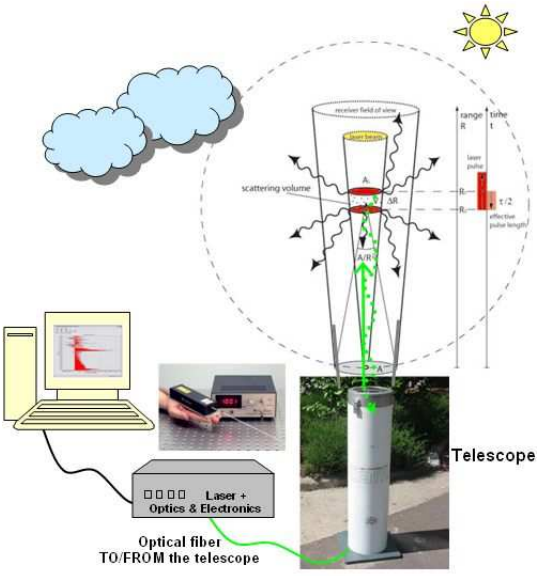

FIGURE 1.4 – Le lidar, combinaison d'un laser de puissance et d'un télescope de précision. [Pappel, 2006]

intertropical (ITD pour Intertropical Discontinuity) et de rayonnement infrarouge sortant au sommet de l'atmosphère (OLR pour Outgoing Longwave Radiation) du modèle de réanalyses NCEP2 ([Hagos and Cook, 2007]).
Ces différents jeux de données illustrent différents problémes qui peuvent être rencontrés dans l'analyse de données géophysiques et relévent tous de la détection et la paramétrisation de signaux cachés. Les observations de PSC présentent des distributions non stationnaires, car le signal contient une tendance non constante, ainsi que la variance (cette derniére propriété est appelée hétéroscedasticité). Les données de carottes de glace, notre premier exemple d'étude multivariée, sont fortement bruitées (avec des variances de bruits différentes pour chaque série de mesure) et ont des tendances également différentes et difficilement paramétrisables. Enfin les données issues du travail sur la Mousson de l'Afrique de l'Ouest présentent des caractéristiques probabilistes similaires à celles des carottes de glace, auxquelles s'ajoutent un échantillonnage irrégulier et un signal caché différent.

3 L'apport de l'analyse probabiliste

L'article de [Zwiers and Von Storch, 2004] explique le réle important du raisonnement statistique dans la recherche climatique, et plus précisément de la prise en compte de l'incertitude et de la variabilité, notamment pour l'acquisition de données, la prédiction météorologique et également dans la compréhension de la dynamique climatique.

L'analyse purement statique de la géophysique s'appuie sur l'étude de fonctions dynamiques déterministes. C'est à dire que pour une méme valeur d'entrée (conditions initiales, conditions limites), un modèle numérique aura toujours la méme valeur de sortie (é erreur-machine prés). Ces études s'appuient en grande partie sur des équations différentielles ordinaires, des équations aux dérivées partielles, la physique statique reléve de l'étude de fonction du type :

$$x_t = f(x_{t-1}, x_{t-2}, ...). \tag{1.1}$$

Cette forme est l'expression discrétisée d'un grand nombre d'équations aux dérivées partielles. Cela suggére que la réalisation x_t à un instant t d'un phénoméne f est fonction exclusivement du passé. En d'autres mots, d'aprés ces équations, la connaissance du passé permet de prédire de maniére sére le futur. Les équations fondamentales de la géophy-

sique (équation de continuité, de mouvement, de conservation d'énergie ..) relévent de cette forme. On trouvera dans les articles de [Malardel, 2005] ou [Amodei and Dedieu, 2008] une introduction des différentes équations fondamentales de la physique ainsi que de leurs analyses numériques.

L'étude des sciences de l'atmosphère par cette approche est limitée à différents niveaux. La premiére limite de cette étude tient du fait que l'ensemble des processus physiques régissant l'atmosphère n'a pas encore été décrit. De plus l'intervention du calcul numérique impose la discrétisation et donc l'approximation des équations continues, qui introduit également de l'erreur. Ainsi, faute de connaissance ou de temps de calcul suffisant, cette approche ne permet pas de prendre en compte certains mécanismes dans l'étude des phénoménes physiques.

Un des réles majeurs de la statistique dans les géosciences est de compléter l'analyse déterministe de la physique. La statistique travaille, par définition sur l'aléa, c'est à dire qu'elle travaille sur la différence entre ce qui est observé et ce que l'on sait expliquer au regard des connaissances établies sur un sujet. Cette information, considérée comme réalisations d'événements aléatoires est modélisée de maniére probabiliste par la fonction de densité de probabilité (fdp) conditionnée par les passées données disponibles :

$$p(x_t|x_{t-1}, x_{t-2}, ...; \theta). \tag{1.2}$$

Dans l'Équation (1.2), le symbole $\cdot|\cdot$ correspond au symbole de conditionnement : l'équation (1.2) représente la densité de probabilité de l'événement x_t, conditionnellement aux événements passés $(x_{t-1}, x_{t-2}, ...)$ et θ représente le vecteurs des paramètres caractéristiques de la fonction de densité p.

Alors que dans l'équation (1.1), il s'agissait de se concentrer sur la réalisation x_t de la fonction f, dans le cas d'une étude probabiliste, l'étude d'un phénoméne repose sur l'analyse de la fdp, qui porte l'information dans le domaine des probables de la répartition des différents événements susceptibles de se réaliser à l'instant t.
Pour de plus amples informations sur les notions de probabilité, on pourra se référer en premiére approche à l'Annexe A, qui reprend les bases nécessaires é la compréhension de ce travail. Pour une étude plus détaillée, de nombreux ouvrages de probabilités existent, une approche simple dans le domaine des sciences de l'atmosphère peut être fournie par

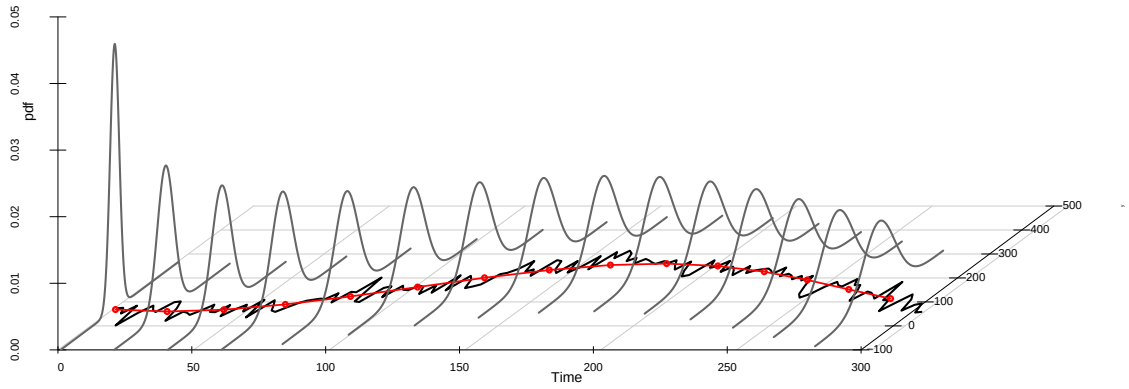

FIGURE 1.5 – La série temporelle bruitée, en noire peut être étudiée soit d'un point de vue déterministe (correspondant à la ligne rouge), soit d'un point de vue probabiliste (par l'étude de l'évolution de la *fdp*)

le livre de [Frontier et al., 2007] et pour une approche plus formelle on consultera des ouvrages plus poussés tel que celui de [Von Storch and Zwiers, 1999].

Cette approche permet ainsi à la fois d'étudier la dynamique d'un systéme et d'en mesurer l'incertitude, et donc notamment de construire un intervalle de confiance bornant de maniére objective le comportement aléatoire du systéme. La figure 1.5 illustre, pour l'étude d'un signal (courbe noire bruitée), la différence de point de vue entre un raisonnement probabiliste (celui de l'étude d'évolution des courbes grises représentant la fdp) et un raisonnement purement déterministe, illustré par les instants de réalisation de la courbe rouge).

Si la forme de la fdp est connue, elle peut parfois être caractérisée par un nombre fini de paramètres. La loi Gaussienne (notée $\mathcal{N}(\mu, \Sigma)$) et définie pour $x \in \mathbb{R}^N$ par :

$$p(x;\theta) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{N/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma(x-\mu)\right), \tag{1.3}$$

avec le vecteur des paramètres caractéristiques $\theta = (\mu, \Sigma)$, où μ représente la moyenne et Σ la matrice de variance-covariance. Dans de tels cas l'étude chronologique de la fdp peut se réduire à l'étude de l'évolution de ses paramètres caractéristiques.

3.1 L'extraction de données

Le premier enseignement fourni par les statistiques est que l'information contenue dans un signal ne doit pas se limiter à l'étude de sa réalisation à un instant donné, que l'ensemble d'un signal apporte de l'information sur sa réalisation à n'importe quel instant t isolé. Formulé encore différemment, le comportement d'un signal à chaque instant peut être caractérisé par l'analyse du signal dans son ensemble. Ainsi par exemple, les moments statistiques (tels que la moyenne et la variance) peuvent être estimés de maniére empirique gréce à l'ensemble des réalisations du signal. Ce comportement statistique peut ne pas être régulier le long du signal, notamment, certains phénoménes peuvent ne pas avoir lieu sur l'ensemble du signal. Les éruptions volcaniques par exemple, n'ont qu'un impact limité dans le temps sur la distribution de sulfate atmosphérique, et ainsi n'altérent le signal de fond que sur une fenétre de temps finie. Pour des raisons évidentes, il est impossible d'observer ou d'étudier le systéme climatique dans son ensemble (ou encore de multiplier les réalisations), et de nombreux phénoménes physiques restent cachés, c'est à dire non directement observés. Il est donc nécessaire et inévitable d'intégrer de l'aléa et de l'incertitude dans les différentes études réalisées.

L'extraction de signaux cachés dans des données a pour but la recherche d'une information précise dont on connait quelques caractéristiques. Ce signal pouvant se cacher dans tout ou partie d'une série temporelle. Dans le cas de l'extraction d'un signal sonore par exemple, on peut imaginer un son d'une fréquence donnée et de forte amplitude auquel s'ajouterait à un instant inconnu un son étouffé de fréquence différente et de plus faible amplitude. Les caractéristiques connues de ces signaux permettent de les isoler l'un par rapport à l'autre. Caractérisés par un certain nombre de paramètres, il devient ainsi possible de rechercher, les événements, porteurs d'informations, que l'ont sait susceptibles de se réaliser dans les données disponibles des sciences de l'atmosphère.
Si le principe est proche de celui de la fouille de données, ou *data mining* ([Tufféry, 2010]), il s'en différencie, tout d'abord dans le sens oé, lors de l'extraction de données, les caractéristiques de l'information recherchée sont supposées connues (dans le chapitre 5, nous modélisons dans le signal des instants de rupture dans la moyenne), ce qui n'est pas le cas dans le *data mining*, dont les méthodes consistent à considérer un ensemble de signaux susceptibles de se réaliser et d'analyser lesquels sont les plus "proches" du signal étudié, en fonction de critéres de distances choisis. De plus, le *data mining* se focalise sur

des méthodes permettant de gérer de jeux de données de dimensions trés grandes.

L'extraction de signaux fait appel à de nombreux champs des statistiques et des probabilités. à partir de la caractérisation probabiliste de phénoménes, elle recherche des occurrences et évalue des incertitudes. Nous revenons au Chapitre 2 plus en détail sur ces différents objectifs fixés.
Ce travail de thèse, à la frontiére entre les probabilités et les sciences de l'atmosphère, explore ce champ de l'extraction de signaux cachés. à travers la recherche et le développement de différents modèles probabilistes nouveaux. Les applications de ces modèles présentées précédemment mettent en lumiére certains phénoménes physiques dans le flot d'informations disponibles, et ainsi fournissent des informations utiles qui ne sont pas directement accessibles. De maniére plus générale, ce travail de thèse illustre par les différentes applications (aux nuages stratosphériques, aux éruptions volcaniques et à la mousson Africaine), le fait qu'il existe des signaux cachés chargés d'information physique exploitable, au prix de l'ouverture vers des nouvelles méthodes probabilistes et du mélange des disciplines.

4 Plan du manuscrit

Les modèles d'inférence probabiliste ainsi que leurs applications à des problémes relatifs aux sciences de l'atmosphère développés lors de ce travail de recherche ont permis la rédaction détaillée de plusieurs articles qui expliquent de maniére exhaustive le travail réalisé, les difficultés rencontrées et le choix des applications. Nous avons ainsi choisi d'insérer directement les articles soumis comme des parties de la thèse.

Le Chapitre 2 introduit la notion d'extraction de signaux cachés, et trace un état de l'art. Les chapitres suivants présentent et développent les articles soumis/publiés pendant cette thèse. Le Chapitre 3 développe un modèle probabiliste de détection d'un signal caractérisé par une rupture transitoire de la variance d'un signal hétéroscedastique. et l'applique à l'étude de détection de nuages stratosphériques polaires en Antarctique. Le Chapitre 4 présente un modèle d'extraction de signaux éruptifs simultanés dans des séries temporelles présentant des tendances différentes et non connues. Ce modèle est appliqué à des séries temporelles de sulfate issues de carottes de glace du Groenland dans le but de détecter des éruptions volcaniques. Le Chapitre 5 se focalise sur un modèle de détection de signaux de rupture dans des séries présentant également des tendances inconnues. Ce

modèle est ensuite appliquée à l'étude de l'onset de la Mousson de l'Afrique de l'ouest, qui est un instant de rupture dans la dynamique atmosphérique continentale. Enfin, le dernier chapitre conclut la thèse et présente certaines perspectives au travail présenté.

Chapitre 2

Rappels sur l'extraction de signaux cachés

Court résumé du chapitre :
Nous présentons tout d'abord dans ce chapitre quelques articles de référence sur l'extraction de données qui ont en partie inspiré cette thèse. Ensuite, nous traitons plus particulièrement des problèmes d'homogénéisation qui alimentent de grands débats dans les sciences de l'atmosphère, et qui font partie des méthodes d'extraction de signaux. Nous présentons ensuite quelques méthodes d'extraction diverses pour finalement introduire le Filtre de Kalman, qui est une approche probabiliste importante dans l'extraction de données et sera, dans le cadre de cette thèse, l'objet de développements aux Chapitres 4 et 5.

Plan du Chapitre 2

1. **Définition de la notion d'extraction de signal**
2. **Rappel sur l'étude des signaux aléatoires non stationnaires**
3. **Les problèmes d'homogénéisation**
4. **Méthodes d'extraction de signaux divers**
5. **Le filtrage de Kalman et l'extraction de signal**

1 Définition de la notion d'extraction de signal

Pour circonscrire ce que sont les méthodes d'extraction de données, il est utile premièrement de faire une distinction entre la statistique descriptive et la statistique inférentielle. La statistique descriptive, également appelée statistique exploratoire est celle ayant pour objectif principal la *représentation* de grands jeux de données. Il s'agit de *représentation* au sens large, l'objectif étant de résumer l'information en un nombre restreint de média, que ce soit par des chiffres ou des graphiques.
La statistique inférentielle, également appelée décisionnelle, quant à elle, a pour objectif de permettre une prise de décision objective au regard de l'information disponible.
C'est dans ce second cadre que se trouvent les méthodes d'extraction de données cachées. La recherche d'une rupture transitoire de la variance d'un signal hétéroscedastique, celle de signaux éruptifs ou encore celle d'instants de variation soudaine de la moyenne dans des séries multivariées non stationnaires posent chacun un problème particulier d'extraction de signal. Le but dans chacun de ces cas étant la recherche et la paramétrisation (timing, amplitude, durée), à partir d'hypothèses sur les caractéristiques statistiques des signaux recherchés. Ces modélisations probabilistes sont enfin confrontées à l'analyse des données, et permettent de détecter ces occurrences et d'évaluer l'incertitude sur ces estimations.

2 Rappel sur l'étude des signaux aléatoires non stationnaires

Les processus stationnaires ont été très tôt étudiés dans les sciences de l'atmosphère. Les travaux de Walker (1868-1958), sont dans ce domaine une référence historique : les équations de Yule-Walker permettent de résoudre le problème de l'estimation des paramètres des modèles Auto-Régressifs, reliant par combinaisons linéaires l'état présent d'un système à ses différents états passés (équation 2.1), et encore largement étudiés ([Katz, 2002]). Un processus Auto-Régressif d'ordre p, $AR(p)$ est défini par :

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \epsilon_t, \tag{2.1}$$

où $\phi_0, ..., \phi_p$ représentent les paramètres du modèle, et ϵ_t un bruit blanc. Ce processus est stationnaire sous la condition que le module des racines de son équation de retard, $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 + .. - \phi_p B^p$, soit strictement superieur à un. Nous utiliserons plusieurs fois de tels modèles dans la suite de ce travail, et plus particulièrement le modèle $AR(1)$, pour *Auto Régressif d'ordre 1* défini par :

$$x_t = \alpha x_{t-1} + \epsilon_t, \tag{2.2}$$

qui est un modèle stationnaire dans le cas où $|\alpha| < 1$, et non stationnaire dans le cas contraire. Nous verrons dans le Chapitre 4 un cas de $AR(1)$ stationnaire, et dans le Chapitre 5 nous verrons les problèmes posés lorsque ce signal n'est pas stationnaire (e.g $\alpha = 1$).

Les sciences de l'atmosphère nécessitent, en effet, l'étude de processus non stationnaires. Dans les années 1980, suite à la parution des articles de [Wahba, 1978] et [Wecker and Ansley, 1983] qui apportent les bases formelles de techniques de traitement du signal appliquées à des signaux non stationnaires, de nombreux travaux apparaissent sur le sujet. Partant du modèle général avec bruit additif :

$$y_i = f(x_i) + \epsilon_i, \tag{2.3}$$

où ϵ_i suit une loi $\mathcal{N}(0, \sigma^2)$, l'article de [Wecker and Ansley, 1983] montre comment mettre en place un modèle de régression non linéaire et non paramétrique pour des signaux non stationnaires basé sur la formulation utilisant des processus auto-régressifs (sans les conditions de stationnarité sur $\phi_0, ..., \phi_p$) définis par l'équation (2.1). Pour modéliser f, Wecker s'inspire du modèle de régression de l'article de [Wahba, 1978] :

$$f(x) = \sum_{k=0}^{m-1} \alpha_k \frac{(x-a)^k}{k!} + \lambda^{1/2}\sigma \int_a^x \frac{(x-h)^{m-1}}{(m-1)!} dW(h), \tag{2.4}$$

où, a est l'instant au voisinage duquel f est approchée, $W(h)$ représente un processus Brownien de variance égale à un, λ représente un paramètre de lissage, $\alpha = (\alpha)_k$ représentent les coefficients de régression polynomiale du modèle et m l'ordre du polynï¿½me de régression utilisé. De manière intuitive, m représente le degré nécessaire de différentiation du signal pour que celui-ci devienne stationnaire. L'article détaille comme résultat,

que toute fonction f dérivable et dont les dérivées jusqu'à l'ordre m sont également dérivables peut être modélisée par l'équation (2.4). D'autres approches sont envisageables pour estimer f, citons notamment une approche par réseaux de neurones (on consultera [Bishop, 2006]), qui représente un compromis entre les approches non paramétrique et paramétrique, dont le principe est basé sur des méthodes d'apprentissage s'appuyant sur des hypothèses quant au comportement de f.

Parmi les études qui suivirent, les articles de [Bell, 1984] et [Kohn and Ansley, 1987] s'intéressent à l'étude de séries de type *signal plus bruit* et permettent grâce à l'étude des espérances et des variances conditionnelles, de décomposer les signaux non-stationnaires de type ARIMA (cf [De Montera, 2008]). Les développements des Chapitres 4 et 5 font directement référence à ces premiers travaux pour estimer la tendance dans des modèles probabilistes d'espace-état.

3 Les problèmes d'homogénéisation

Parmi les différentes approches de la recherche de signaux cachés, les méthodes d'homogénéisation ont une place particulière à la fois parmi les méthodes d'extraction et dans les sciences de l'atmosphère. Ces techniques recherchent dans des séries temporelles des changements artificiels abrupts. Ces perturbations peuvent tout autant porter sur la moyenne, sur la variance, ou plus généralement sur la distribution de probabilité des données. Une fois détectées, ces ruptures artificielles sont corrigées afin de rendre la série homogène.

Ces problèmes ont toujours une forte actualité, s'il en est pour preuve les récents discours antagonistes entre d'une part Jean-Louis Mouï¿½l et Vincent Courtillot et d'autre part Edouard Bard, Bernard Legras, Pascal You et Olivier Mestre, lors de la séance de débats organisée à l'Académie de Sciences à Paris en septembre 2010, qui a donné lieu à de vifs échanges. Le groupe dï¿½E.Bard reprochant à celui de J.L. Mouï¿½t d'avoir étudié des données non homogénéisées pour servir leur argumentation. Le compte rendu succinct publié par l'Académie des Sciences en Octobre 2010 ([Puget and Blanchete, 2010]) fait également plusieurs fois mention du problème de l'homogénéisation des données, et conclut entre autre que "*l'évolution du climat ne peut être analysée que par de longues série de données, à grande échelle, homogènes et continues*", point de vue soutenu dans différents travaux antérieurs tels que [Legras et al., 2010], [Mestre and Caussinus, 2001] et [Moisselin et al., 2002].
Plus largement, d'autres domaines s'intéressent à ces problématiques. La finance, les pro-

blèmes de fiabilité ou le contrï¿½le de production par exemple (e.g. [Egon and Porée, 2003]) utilisent des méthodes similaires à celles illustrées précédemment.

Cette procédure d'homogénéisation est le préalable à une étude plus approfondie des données. Dans des contextes de longues séries de données, comme par exemple des mesures datant de plusieurs décennies, il est concevable que les instruments de mesures aient été entretenus, rénovés, déplacés, voire même la technique de mesure modifiée, ce type de changements pouvant affecter les mesures. De nombreuses recherches se sont intéressées à ces problèmes, les prémices historiques de ces méthodes étaient basées sur la mise en place de tests statistiques tels que le test de Student pour des ruptures de moyennes (e.g. [Gosset, 1908]). En nous plaïï¿½ant dans le cadre de deux échantillons x et y, dont ni les moyennes ni les variances sont connues, mais en considérant que les variances sont égales, le test de Student est défini par :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2(\frac{1}{n_x} + \frac{1}{n_y})}}, \tag{2.5}$$

où $\bar{x}$ et $\bar{y}$ sont les moyennes empiriques de deux échantillons x et y et s^2 la variance empirique des échantillons x et y.
L'estimateur ainsi défini suit une distribution *t de Student*, $t_{n_x+n_y-2}$. Le calcul de t ainsi que du degré de liberté associé à cette estimation seront utilisés sur une table de *distribution t* pour tester l'hypothèse nulle d'égalité des moyennes, $\bar{x}$ et $\bar{y}$, des deux échantillons.
Un second test historique est celui de Fisher (e.g. [Fisher-Box, 1987]) qui permet de tester l'égalité entre les variances de deux échantillons indépendants et qui consiste à calculer le rapport défini par :

$$F = \frac{\frac{n_x}{n_x-1}s_x^2}{\frac{n_y}{n_y-1}s_y^2}, \tag{2.6}$$

avec comme hypothèse nulle l'égalité des variances. Cette quantité F suit une loi de Fisher-Snédécor $\mathcal{F}(n_x - 1, n_y - 1)$; elle est calculée comme le rapport de deux variables aléatoires s_x^2 et s_y^2 qui suivent chacune une loi du *Chi-Deux* (carrés de variable gaussiennes). Le test revient à comparer ce résultat à la table de Fisher $\mathcal{F}(n_x - 1, n_y - 1)$, qui indique l'égalité ou non des variances en fonction d'un seuil de significativité fixé α. On rejettera l'hypothèse d'égalité si la valeur de F est trop grande ou trop petite. Ces premiers estimateurs, basés sur une approche fréquentiste sont très limités. Tout d'abord,

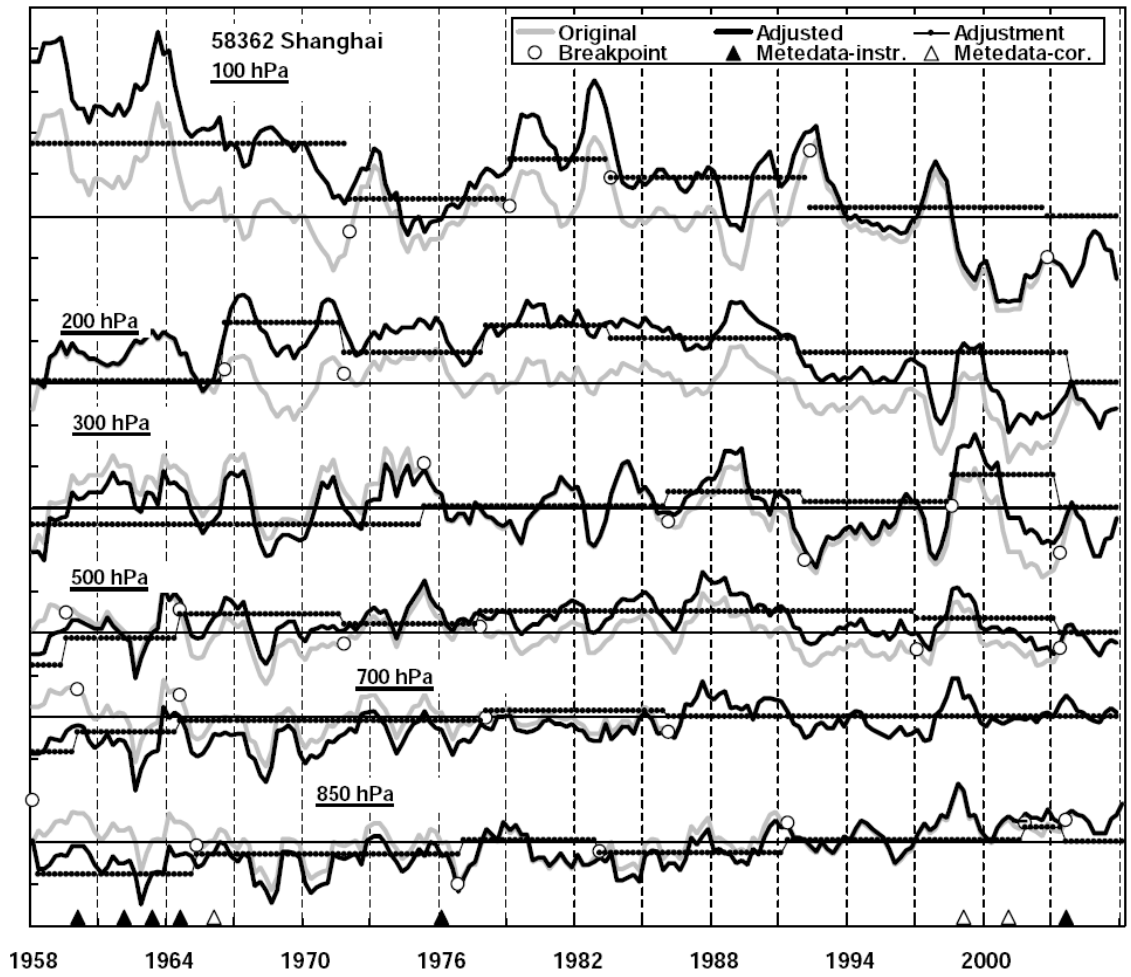

FIGURE 2.1 – Procédure d'homogénéisation de séries de températures issues de l'article de [Guo and Ding, 2009] – **Légende initiale :** *Four-point running-mean seasonal temperature anomalies (K) at station 58362 for 100- 850 hPa. Filled circles indicate detected breakpoints, step-function curves denote adjustments, and black and gray lines indicate homogenized and original radiosonde temperature time series. Dynamic metadata events are indicated by solid (or dashed) vertical lines for instrument (correction method) changes. Each interval on the vertical axis indicates 1 K.*

ils supposent que les échantillons x et y soient indépendants et qu'ils suivent des lois Gaussiennes. Appliqués à une série temporelle, ils supposent connu l'instant de *partage* de la série, à savoir, le moment où la distribution de probabilité des échantillons change (changement de moyenne ou de variance). Ces hypothèses parfois difficiles à vérifier dans la pratique, ne permettent pas de traiter les cas souvent compliqués rencontrés dans les sciences de l'atmosphère, et notamment, ils ne permettent la détection que d'une seule rupture par test. Nous utilisons notamment le test de Fisher-Snédécor, combiné à une méthode de stationnarisation et à un calcul d'un rapport de vraisemblance, dans le Chapitre 3.

Parmi les travaux plus récents, l'article de [Davis et al., 2006] développe des méthodes basées sur des calculs de maximum de vraisemblances pénalisées par un critère de MDL (Maximum Description Length). Cette méthode paramétrique permet de s'affranchir des certaines hypothèses, telle que le caractère Gaussien de la distribution étudiée, et permet de détecter plusieurs ruptures dans une même série. La méthode de [Davis et al., 2006] suppose cependant que les différentes partitions de la série suivent des modèles Auto-Régressifs. Par principe, l'approche par maximisation de vraisemblance pénalisée utilisée par Davis et al. [2006] revient premièrement à considérer la fonction de vraisemblance :

$$\mathcal{L}(\theta) = p(x; \theta), \tag{2.7}$$

où $x = (x_1, \ldots, x_n) \in \mathbb{R}^N$ représente un échantillon de taille N, et θ les paramètres de la densité p. La fonction $\mathcal{L}$ à la même forme que la densité de probabilité définie à l'équation (1.2), mais est considérée comme une fonction dont les variables sont les paramètres θ et non plus les réalisations x, le vecteur x étant considéré constant.
La maximisation de la vraisemblance pénalisée cherche le minimum donné par :

$$\arg\min_{\theta \in \Theta} \{-\log \mathcal{L}(\theta) + w(\theta)\}, \tag{2.8}$$

où Θ représente l'ensemble des paramètres, et la fonction $w(\theta)$ est un critère qui tend à ajouter du poids à certains paramètres. Par exemple les critères AIC (pour Akaike Information Criterion, voir [Akaike, 1974]), BIC (pour Bayesian Information Criterion, voir [Schwarz, 1978]), ou MDL (pour Minimum Description Length, voir [Rissanen, 1978]) tendent à ajouter du poids au nombre de paramètres par des approches différentes, et ainsi à favoriser les vraisemblances comportant le moins de paramètres. Une fois la fonction

de vraisemblance et la pénalisation choisies, différentes méthodes existent pour opérer la maximisation. Si le calcul formel est possible, cette maximisation est résolue par l'étude de la fonction à plusieurs variables $\mathcal{L}$, ceci est possible sous certaines conditions de continuité qui ne sont pas détaillées dans ce manuscrit. Lorsque le calcul formel est rendu impossible, des solutions approximées peuvent être calculées par des algorithmes tels que l'algorithme EM (pour Expectation-Maximization, voir [Dempster et al., 1977]) et ses dérivés, ou encore des méthodes de Monte-Carlo. La maximisation d'un rapport de vraisemblance sera l'objet du développement du modèle présenté au Chapitre 3

Le livre de [Basseville and Nikiforov, 1996] explore de nombreuses méthodes de détection de changements dans des séries temporelles. Ce livre de référence pose des problèmes théoriques généraux de détections et une large gamme de situations, tels que des problèmes univariés et multivariés et des problèmes de détection en temps réel.
Les problèmes posés par les sciences de l'atmosphère étant assez spécifiques, des travaux plus récents développent des méthodes plus spécialisées. Certaines études utilisent par exemple des séries de référence en comparaison de la série à étudier : [Staehelin et al., 2009] utilisent des séries de températures comme référence pour tester l'homogénéité de longues séries d'ozone. D'autres utilisent des méthodes de régression combinées à des tests statistiques : l'article de [Guo and Ding, 2009] (voir figure 2.1), utilisant également des méta-données, recherche dans des séries de températures un instant de rupture de modèle de régression linéaire correspondant à une datation du changement climatique. D'autres enfin, grâce à des méthodes basées sur une maximisation de vraisemblance pénalisée étudient des séries temporelles climatiques et extraient des ruptures multiples significatives permettant d'homogénéiser ou de détecter des changements climatiques. Parmi ces méthodes, l'article de [Caussinus and Mestre, 2004] met en place un calcul de maximisation de vraisemblance en pénalisant les modèles notamment en fonction de leur nombre de paramètres et du nombre de données, ceci sur des séries de températures. Enfin, [Hannart and Naveau, 2009] se sont également intéressés à des problèmes d'homogénéisation, utilisant des méthodes bayésiennes sur des séries climatiques. [Reeves et al., 2007] résument et comparent quelques méthodes de détection d'instants de rupture. Depuis quelques décennies donc, parmi d'autres problèmes d'extraction, ceux touchant à l'homogénéisation sont largement débattus dans la communauté des géoscientifiques.

4 Méthodes d'extraction de signaux divers

L'extraction de données au sens plus large touche des problèmes plus vastes. Nous citons dans cette section quelques exemples qui seront pris en référence dans la suite du manuscrit ou certains autres traitant spécifiquement des problèmes des sciences de l'atmosphère.

L'article de [Guo et al., 1998] est plusieurs fois cité dans ce manuscrit. Ce travail met en place une technique univariée de détection de signaux éruptifs dans des données d'hormones, permettant, à partir des séries d'hormones, de détecter les instants d'injections et les vitesses de décroissance des concentrations dans les échantillons. Les chapitres 4 et 5 s'appuient sur ce travail poussant le développement de la méthode à un cadre multivarié (Chapitre 4) ou adaptant la base de ce travail pour aboutir à une méthode de détection de rupture là aussi dans un contexte multivarié (Chapitre 5).
Dans le domaine climatique, différents travaux, encouragés par les recherches du Panel Intergouvernemental sur le Changement Climatique (en anglais Intergovernmental Panel on Climate Change - IPCC) se sont focalisés sur les marqueurs du changement climatique. L'idée principale de la démarche menée par l'IPCC étant de différencier la détection des changements (e.g. [Ribes et al., 2010]), et l'attribution de ces changements (e.g. [Hegerl et al., 2007]). De manière schématique, dans cette distinction, la détection correspond à l'étape de caractérisation des changements, en temps, intensité, localisation ; l'attribution examine a posteriori les causes de ces changements. Plus récemment, les travaux de [Boreux et al., 2009], utilisant une méthode bayésienne, ont identifié des variations hautes fréquences communes dans la croissance des arbres d'une même région à partir de cernes de conifères de la région nord du Québec, Canada. Cette étude multivariée, c'est à dire l'analyse simultanée à plusieurs séries temporelles, permet d'extraire des signaux communs à chacun des arbres.

De manière plus générale, l'extraction de signaux cachés peut aussi être appréhendée en termes d'étude de variables latentes. Le chapitre 4 de [Droesbeke et al., 2005] écrit par Gilbert Saporta revient sur cette notion. L'idée principale étant la recherche de variables qui ne sont pas directement observables mais qui s'expliquent par leurs dépendances avec les variables observées. Le tableau 2.1 résume les différents modèles d'étude des variables latentes en fonction du type de variables recherchées et observées.
La recherche de variable latente est plus générale que l'extraction de signaux cachés, dans le sens où elle n'est pas circonscrite à l'étude de signaux à partir de séries de don-

TABLE 2.1 – Modèles à classes latentes

	Variables latentes	
Variables observées	qualitatives	quantitatives
qualitatives	Analyse des classes latentes	Analyse des traits latents
quantitatives	Analyse des profiles latents	Analyse factorielle

nées. Elle recherche des variables tant quantitatives que qualitatives à partir de jeux de données non nécessairement ordonnés, comme le sont les séries. En ce sens, l'approche par variables latentes regroupe plusieurs méthodes d'inférence, telles que l'utilisation de chaîne de Markov, l'analyse en composantes principales, et plus généralement l'analyse factorielle. Pour plus de détails on consultera les ouvrages de [McCutcheon, 1987] ou encore [Droesbeke et al., 2005].

5 Le filtrage de Kalman et l'extraction de signal

Dans cette section, nous rappelons quelques résultats sur le filtrage et le lissage issus du système d'état/observation de Kalman. Ces résultats sont à la base des travaux menés aux Chapitres 4 et 5.
En traitement du signal, un filtre $\mathcal{F}$ est un opérateur qui transforme un signal $x \in \mathbb{R}^n$ en un autre signal $y \in \mathbb{R}^m$:

$$y = \mathcal{F}(x) + \epsilon, \tag{2.9}$$

où ϵ est une variable aléatoire généralement centrée (un bruit blanc par exemple). Lorsque la transformation par $\mathcal{F}$ est linéaire et le bruit ϵ est gaussien, on parle de filtre linéaire. Parmi les exemples de filtres, la transformée de Fourier et les moyennes glissantes sont fréquemment utilisées dans les sciences de l'atmosphère, respectivement pour déplacer le signal initial dans le domaine fréquentiel, et pour effectuer un lissage de données (supprimer les composantes hautes fréquences du signal). Le problème du filtrage consiste à considérer un processus $x \in \mathbb{R}^n$, dont on connait les caractéristiques statistiques et qui représente l'état d'un système non observé ou partiellement observé. Ce qui est observé est le processus $y \in \mathbb{R}^m$ auquel s'adjoint un bruit additif ϵ, dont les caractéristiques sont également connues. A chaque instant, l'information collectée est donnée par le vecteur de observation : $Y_t = (y_1, \ldots, y_t)$. Le but du filtrage étant de calculer à chaque instant t une estimation de x au regard de l'information disponible Y_t.

Les filtres font partie des méthodes d'extraction d'information, dans le sens où, si on considère un état x composé d'éléments dynamiques et d'éléments cachés, le filtrage aura pour objectif de différencier ces différentes composantes l'une de l'autre.

Le filtre de Kalman (KF) (e.g. [Kalman, 1960], [Kalman and Bucy, 1961]) est une méthode probabiliste d'assimilation de données, son principe consiste à suivre un modèle, $x \in \mathbb{R}^n$, défini par une équation d'état, et d'en corriger les trajectoires (x_t) à l'instant t grâce à des données d'observation Y_t défini par l'équation d'observation. Il s'agit donc d'ajouter à l'équation d'observation (2.9), une équation d'état. Soit l'équation d'observation notée :

$$x_t = \Phi x_{t-1} + \epsilon_t^*, \tag{2.10}$$

et l'équation d'observation sera re-notée, dans le cas où $\mathcal{F}$ est linéaire :

$$y_t = H x_t + \epsilon_t, \tag{2.11}$$

où Φ et H sont des opérateurs linéaires, $\epsilon_t \in \mathbb{R}^m$ et $\epsilon_t^* \in \mathbb{R}^n$ représentent respectivement le bruit d'observation et l'erreur du modèle.

Les équations du filtre de Kalman, sous les hypothèses de linéarité des équations d'état et d'observation (équations définies par le système (2.10-2.11)) et sous l'hypothèse de bruits gaussiens se résolvent par l'estimation notée $\hat{x}_t \in \mathbb{R}^n$ de la trajectoire aux différents instants. Cela se fait par minimisation de l'erreur quadratique moyenne, ou encore par minimisation de la variance de l'erreur d'estimation de l'état du système au regard des observations. Cette minimisation est équivalente, sous les hypothèses citées précédemment de linéarités, à calculer la distribution conditionnelle $p(x_t|Y_t)$ introduite par l'équation (1.2). Sous ces hypothèses, la distribution $p(x_t|Y_t)$ restant Gaussienne, la trajectoire $\hat{x}_t$ est entièrement caractérisée par ses moments d'ordre un et deux, à savoir, sa moyenne et sa matrice de variance-covariance :

$$\hat{x}_t = \mathbb{E}[x_t|Y_t], \tag{2.12}$$

$$\hat{\Sigma}_t = \mathbb{E}[(x_t - \hat{x}_t)(x_t - \hat{x}_t)'|Y_t]. \tag{2.13}$$

L'article de [Kalman and Bucy, 1961] présente l'algorithme de Kalman-Bucy : il s'agit d'un algorithme récursif, qui nécessite à chaque étape un calcul de *prédiction* et un calcul de *correction*.
La *prédiction* correspond au calcul de la loi $p(x_t|Y_{t-1})$ c'est à dire, la densité de x_t connaissant les observations jusqu'à l'instant $t-1$, notées Y_{t-1}.
Cette étape est corrigée par le calcul de *correction*, qui consiste à utiliser l'information nouvelle apportée par la dernière observation y_t par rapport à l'information connue Y_{t-1} lors de l'étape précédente. Cette information est contenue dans l'*innovation*, définie par : $I_t = y_t - \mathbb{E}[y_t|Y_{t-1}] = y_t - H\mathbb{E}[x_t|Y_{t-1}]$. La *correction* correspond à l'étape qui calcule l'erreur commise entre l'estimation de x_t connaissant les observations jusqu'au temps t, et l'estimation de x_t connaissant les observations jusqu'au temps $t-1$ (cette dernière estimation correspondant au calcul de *prédiction*). Ainsi grâce aux calculs suivant :

$$\mathbb{E}[x_t|Y_{t-1}] = \Phi\hat{x}_{t-1}, \tag{2.14}$$
$$\hat{\Sigma}[x_t|Y_{t-1}] = \Phi\hat{\Sigma}_{t-1}\Phi' + \Sigma_{\epsilon^*}, \tag{2.15}$$

où Σ_{ϵ^*} est la matrice de variance-covariance de ϵ_t^*.
On obtient alors l'estimation du filtre de Kalman via les formules :

$$\hat{x}_t = \mathbb{E}[x_t|Y_{t-1}] + K(y_t - H\mathbb{E}[x_t|Y_{t-1}]), \tag{2.16}$$
$$\hat{\Sigma}_t = (I - KH)\hat{\Sigma}[x_t|Y_{t-1}], \tag{2.17}$$

où K représente la matrice du gain de Kalman et est donné par :

$$K = \hat{\Sigma}[x_t|Y_{t-1}]H'[H\hat{\Sigma}[x_t|Y_{t-1}]H' + \Sigma_\epsilon]^{-1}, \tag{2.18}$$

où Σ_ϵ, la matrice de variance-covariance de e_t est supposée inversible.
Une dernière étape de résolution tient dans le *lissage* des données, qui est la dernière étape de la décomposition d'un signal. Il s'agit, une fois le filtrage de Kalman réalisé, de reconstruire l'ensemble des trajectoires, non plus au regard de l'information disponible au temps t, mais au regard de l'ensemble des informations disponibles, à savoir, l'ensemble des observations $Y_n = (y_1 \ldots, y_n)$. Cette étape sera nommée également FIS dans la suite de ce travail, pour *Fixed Interval Smoother* (voir Annexe B). Cette étape consiste donc à

calculer, à partir des calculs précédents les projections $\mathbb{E}[x_t|Y_n]$ et $\Sigma[x_t|Y_n]$ définies par :

$$\mathbb{E}[x_t|Y_n] = \hat{x}_t + C[\mathbb{E}[x_{t+1}|Y_n] - \Phi\hat{x}_t], \tag{2.19}$$
$$\Sigma[x_t|Y_n] = \hat{\Sigma}_t + C[\Sigma[x_{t+1}|Y_n] - \Sigma[x_{t+1}|Y_t]]C', \tag{2.20}$$

où la matrice $C = \hat{\Sigma}_t\Phi[\Sigma[x_{t+1}|Y_t]]^{-1}$ et $\Sigma[x_{t+1}|Y_t] = \Phi\hat{\Sigma}_t\Phi + \Sigma_{\epsilon*}$.
Il faut noter que ces dernières relations de récurrence se font *en remontant le temps* : la récurrence se fait de $t+1$ à t. Pour plus de détails sur les calculs précédents, on pourra consulter l'appendice A, ou des articles tels que [Sarkka et al., 2006] ou [Evensen, 2009].

Depuis les premières publications dans les années 60, de nombreux développement ont été apportés. Or le système (2.10-2.11) du filtre de Kalman est limité par la linéarité des deux équations et il est nécessaire de modéliser des processus non-linéaires. Pour cela ont été développés des filtres de Kalman prenant en compte de la non linéarité, autant dans l'équation d'état que d'observation. Le filtre de Kalman étendu (EKF), (e.g. [Welch and Bishop, 2006]), par exemple, généralise le système (2.10-2.11) au système :

$$\begin{aligned} x_t &= \phi(x_{t-1}) + \epsilon_t^*, \\ y_t &= h(x_t) + \epsilon_t, \end{aligned} \tag{2.21}$$

où ϕ et h sont non linéaires et ϵ_t^* et ϵ_t sont respectivement les bruits de modèle et d'observation. Dans ce cadre, la résolution du système défini par (2.21) du filtre de Kalman étendu s'opère par linéarisation des fonctions ϕ et h autour de $\hat{x}_{t-1}$:

$$\phi(x) \simeq \phi(\hat{x}_{t-1}) + \phi'(\hat{x}_{t-1})(x - \hat{x}_{t-1}), \tag{2.22}$$
$$h(x) \simeq h(\hat{x}_{t-1}) + h'(\hat{x}_{t-1})(x - \hat{x}_{t-1}), \tag{2.23}$$

qui est une approximation du développement de Taylor à l'ordre 1.
Dans les Chapitres 4 et 5, nous développons des filtres de Kalman dont les équations d'état et d'observation sont proches du modèle (2.21), à cela près que les fonctions ϕ et h y sont considérées inconnues et que des signaux cachés sont ajoutés à l'équation d'observation (2.21).

L'utilisation du filtre de Kalman est très étendue et se retrouve dans divers domaines. On peut citer entre autres exemples, la balistique pour le calcul de trajectoires ([Villien, 2006]), l'hydrologie, pour des problèmes de prévisions de débits ([Ouachani et al., 2010]), la finance ([Manoliu and Stathis, 2002]) pour l'estimation et la prévision de la volatilité stochastique, ou encore la géolocalisation par système de positionnement satellitaire. Les sciences de l'atmosphère aussi regorgent d'exemples d'utilisations du filtre de Kalman, l'utilisation la plus répandue étant l'assimilation de données, technique qui consiste à utiliser des données d'observation pour affiner le calcul, fait à partir de modèles numériques, de systèmes tels que l'état de l'atmosphère, la température de l'océan, etc .. voir [Talagrand, 2003].

Dans les différentes études présentées dans cette thèse, le filtre de Kalman a été utilisé pour le traitement de signaux aléatoires en ayant comme objectif la décomposition de signaux additifs. De telles méthodes cherchent à distinguer les différentes composantes d'un signal dont certaines se réalisent à des temps aléatoires inconnus. Il faut en extraire de manière distincte les différentes composantes, en vue de pouvoir les étudier de manière indépendante, et d'estimer l'amplitude et la fréquence d'occurrences des phénomènes cachés. Dans les Chapitres 4 et 5, il s'agit d'extraire des signaux particuliers, dont les caractéristiques sont définies à partir d'hypothèses quant à la dynamique du phénomène étudié. Plus précisément, pour le Chapitre 4 (respectivement le Chapitre 5), le modèle permet d'identifier des empreintes que laissent les éruptions volcaniques dans des carottes de glace (resp. des phénomènes de ruptures illustrant la mousson africaine). Les problèmes posés aux Chapitres 4 et 5, nous ont ainsi amené à résoudre les équations du filtre de Kalman dans deux cas non stationnaires et pour lesquels des non linéarités apparaissent à des instants aléatoires inconnus le long des trajectoires.

Une autre méthode de résolution du système (2.21) est celui du filtre particulaire (Doucet et al. [2001]). Il s'agit d'implémenter par des méthodes de Monte-Carlo, un filtre bayésien récursif. Il permet de traiter des systèmes non linéaires dont on ne peut calculer la solution analytique. Il consiste à observer le comportement d'un jeu de particules dont les mouvements sont soumis aux équations du système de Kalman défini par les équations (2.21) (le mouvement d'une particule représente une trajectoire). À chaque pas de temps du système, la loi des particules est échantillonnée et les particules sont regroupées et pondérées en fonction de l'évolution de cette loi afin de contrï¿½ler le nombre de particules et ainsi d'alléger les calculs suivants.
Enfin la résolution du système (2.21) est rendue possible dans le cadre de système de grande dimension, grâce au développement du filtre de Kalman d'ensemble (e.g. [Even-

sen, 2003], [Evensen, 2009]), également basé sur des méthodes de Monte-Carlo, et qui permet d'estimer de manière empirique et simplifiée la matrice de covariance d'erreur, qu'il est impossible de calculer dans la pratique lorsque le système est de trop grande dimension.

Chapitre 3

Détection de rupture transitoire de variance : application à la détection et à la paramétrisation des nuages stratosphériques polaires

Court résumé du chapitre :
Ce premier chapitre présente une application liée à la détection de nuages stratosphériques polaires au travers du modèle probabiliste suivant :

$$P(z) = m(z) + x(z) + \sigma^2(z)\epsilon(z). \tag{3.1}$$

Ce modèle a pour objet la détection d'une rupture transitoire de variance, à savoir, la détection dans une série de données, de deux altitudes (dans cette application les séries sont ordonnées par l'altitude) entre lesquelles la variance de la série augmente (la Figure 3.1 illustre un tel signal). Le modèle additif étudié est composé d'une tendance (m), d'un signal (x) suivant une distribution gaussienne de moyenne nulle ($\mathcal{N}(0, \sigma_c^2)$), égale à zéro à l'extérieur d'un intervalle inconnu et d'un bruit héteroscedastique $\sigma^2(z)\epsilon(z)$. Ce développement est basé sur la maximisation d'une vraisemblance et la mise en place de tests d'hypothèses. Il est appliqué à la détection de nuages stratosphériques polaires en Antarctique. Nous introduisons cette étude grâce à un préambule, suivi par un article qui sera prochainement soumis.

Plan du Chapitre 2

1. **Préambule**
2. **Introduction**
3. **Lidar data**
4. **An procedure to detect PSCs**
5. **The effect of temporal averaging of profiles using real data**
6. **Discussion and conclusion**
7. **Appendices**

1 Préambule

Les nuages stratosphériques polaires (PSC) sont le lieu de nombreuses réactions chimiques au centre de la destruction de l'ozone polaire et ainsi sont au centre des processus de formation du trou de la couche d'ozone (voir [WMO, 2007] et [Peter, 1997]). Leur détection permet d'étudier leur processus de formation et le lien entre leur présence et la distribution de l'ozone stratosphérique polaire. Un second intérêt à détecter ces nuages à partir des données brutes de rétrodiffusion lidar est que cela permet d'adapter a priori les paramètres des équations d'inversion lidar à l'origine des profils de rapport de diffusion plus communément étudiés dans la littérature ([David et al., 2009]). Les signaux bruts de rétro diffusion donnant la puissance recue ($W.m^{-2}$) mesurée par le télescope (voir Figure 1.4) sont donnés par la formule :

$$F_0\beta(z)\frac{K}{z^2}\exp[-2\int_{z_0}^{z}\alpha(z')dz'], \tag{3.2}$$

où F_0 est l'énergie initiale en sortie du lidar, β est la somme du coefficient de rétrodiffusion moléculaire et du particulaire (aérosols, nuages), K est une constante instrumentale (prenant en compte la surface du récepteur), z_0 est l'altitude de l'instrument, α est le coefficient d'extinction, et z l'altitude (m) supérieure à z_0. Cette équation est utilisée pour remonter au coefficient de rétrodiffusion à partir du signal brut P et de la sensibilité K du système. Pour plus de détails, on consultera [David, 1995], [Collis and Russell, 1976], [Fierli et al., 2001] et [David et al., 2004]).
Ce coefficient est ensuite utilisé pour calculer le rapport de diffusion qui est la grandeur de référence pour l'étude des données lidar et qui est exprimé par le rapport :

$$R(z) = \frac{\beta_\alpha(z) + \beta_m(z)}{\beta_m(z)}, \tag{3.3}$$

où β_m est le coefficient de retrodiffusion moléculaire et β_α le coefficient de rétrodiffusion particulaire. Dans ces équations, la présence de nuages stratosphériques polaires intervient dans les paramètres β et α de l'équation (3.2), qui est présenté ici de manière très simplifiée. Le coefficient α décrit l'extinction du faisceau laser à la fois lors de sa montée vers la cible et lors de sa descente vers les capteurs du télescope. On pourra en trouver une formulation plus détaillée dans de nombreux ouvrages tels que celui de [Kovalev and

Eichinger, 2004].

Ce chapitre présente un cas d'étude de détection de rupture transitoire de la variance d'un signal, ici due à la présence d'une couche de PSC. Par rupture transitoire de variance, il est entendu une augmentation soudaine et localisée de la variance entre deux altitudes inconnues du signal. Ce travail présente plusieurs particularités, tout d'abord, les données étudiées sont des profils verticaux de rétrodiffusion issus de mesures lidar effectuées en Antarctique. Il n'est pas ici question de signaux temporels, les variations du signal rétrodiffusé sont étudiées le long d'un axe d'altitude. Ensuite, il s'agit d'une étude dans laquelle le signal étudié n'est initialement pas stationnaire : le signal présente une tendance et sa variance n'est pas constante (caractéristique appelée hétéroscedasticité). Même lorsque le signal (profil lidar de rétrodiffusion) ne contient pas de couche de PSC, la variance évolue avec l'altitude. Deux difficultés principales ont été rencontrées. Tout d'abord comment stationnariser, au sens probabiliste, des profils de rétrodiffusion, ensuite quelle méthode de détection mettre en place afin d'inférer sur la présence de nuage dans un profil et d'en extraire les domaines d'altitudes. Il a donc été nécessaire d'analyser différentes techniques de stationnarisation avant de mettre en place une méthode de détection du nuage.

Dans le modèle choisi défini par l'équation (3.1), $m(z)$ modélise la tendance non linéaire correspondant à la distribution d'aérosols en fonction de l'altitude, décrit par [Junge et al., 1961], qui peut être considérée comme la quantité d'équilibre des aérosols stratosphériques, à savoir la répartition des aérosols lorsqu'il n'y a ni nuage, ni éruption volc anique. Le signal $x(z)$ correspond au signal caché, un signal nul sauf dans un intervalle, dans lequel il reste de moyenne nulle, mais présente une forte variabilité (notée σ_{cloud}). La Figure 3.1 présente une illustration d'un tel signal. Enfin, le signal $\sigma^2(z)\epsilon(z)$ est un signal de moyenne nulle également, mais hétéroscédastique, dont la variance évolue selon $\sigma(z)$, et qui correspond au bruit de fond du signal rétrodiffusé. (Ce modèle est présenté plus schématiquement dans l'article par : $P_{raw} = P_{trend} + P_{cloud} + P_{back}$.) Afin de résoudre la décomposition de ce modèle, nous avons développé une méthode de stationnarisation et un calcul de maximisation de vraisemblance. Cette approche a pour but de déterminer quel modèle (celui avec ou celui sans PSC) correspond le plus vraisemblablement au profil étudié. Il s'agit de déterminer quelles seraient les altitudes d'un profil les plus probables à contenir un PSC, et ensuite de statuer sur l'égalité ou non des variances de deux signaux obtenus après différentes opérations sur le signal initial.

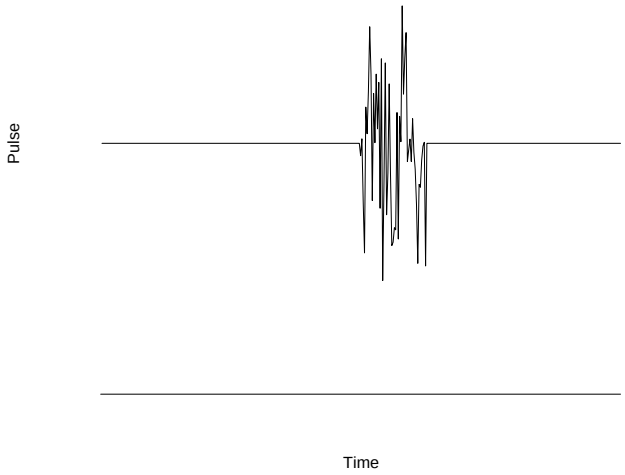

FIGURE 3.1 – Une simulation du signal x recherché de rupture transitoire de variance de l'équation (3.1).

Nous montrons notamment dans cette étude l'impact qu'il y a d'utiliser des profils moyennés dans l'étude des Nuages Stratosphériques Polaires. Il s'avère que de moyenner les profils sur plus de deux heures diminue sensiblement la qualité des profils, certains nuages ne sont plus détectés, et d'autres disparaissent dans les profils lors du processus de moyennage.
Les résultats positifs de cette étude permettent d'envisager des applications plus larges telles que le développement de ce modèle de détection d'une couche nuageuse en un modèle de détection multicouches (de plusieurs nuages) apparaissant dans un même profil. Cela revient à considérer que x contient plusieurs ruptures transitoires d'augmentation de variance. Il est également envisagé d'autres applications telles que l'adaptation de ce modèle à la détection de panaches de cendres ou de poussières issus d'éruptions volcaniques. Parmi les perspectives d'amélioration, nous avons également envisagé la prise en compte de la dimension temporelle et saisonnière. Le modèle (3.1) proposé ne considère pas l'aspect saisonnier de l'apparition des nuages, ni la possibilité qu'un même nuage peut apparaître sur plusieurs profils successifs. Ces considérations pouvant être modélisées par une dépendance temporelle de la détection.

Après une introduction sur les problèmes posés par les nuages stratosphériques polaires et leurs détections et une section sur les caractéristiques des données lidar, nous

expliquons le modèle de détection mis en place, à savoir la stationnarisation des profils, le calcul de paramètres utiles pour établir la vraisemblance, puis l'approche récursive choisie pour le calcul du maximum. Enfin, nous détaillons les résultats sur les données simulées ainsi que sur les profils de la station de Dumont d'Urville, Antarctique.

Article #1 : doi :10.5194/acp-12-3205-2012 - ***Atmospheric Chemistry and Physics***

Detection of Polar Stratospheric Clouds in backscatter profiles over Dumont D'Urville, Antarctica

Gazeaux Julien[1], Bekki Slimane[1], Naveau Philippe[2], Jumelet Julien[1], Keckhut Philippe[1], Parades Jose[1], and David Christine[1]

[1]UPMC Univ. Paris 06 ; Université Versailles St-Quentin ; CNRS/INSU, LATMOS-IPSL, France, Paris

[2]Laboratoire des Sciences du Climat et de l'Environnement (LSCE), IPSL, CNRS/CEA, France, Gif-Sur-Yvette

abstract

Detection methods is proposed and studied to infer the presence of hidden signals in a statistical way. It is applied here to the detection of Polar Stratospheric Clouds (PSC) layers in lidar backscatter profiles measured over the Dumont D'Urville station (Antarctica). PSCs appear as layers with enhanced variance in non stationary, heteroscedastic signal profiles, between two unknown altitudes to be estimated. The method is based on a three step algorithm. The first step is the stationarization of the signal, the second performs the maximum likelihoods estimation of the signal (PSC altitude range and variance inside and outside the PSC layer). The last step uses a Fisher-Snédécor test to decide whether the detection of PSC layer is statistically significant. Performances and robustness of the method are successfully tested on simulated data with given statistical properties. Bias and detection limit are estimated. The method is then applied to lidar backscatter profiles measured in 2008. No PSC are detected during seasons when PSCs are not expected to form. As expected PSC layers are detected during the austral winter and early spring. The effect of time averaging of the profiles is investigated. The result suggests that the best compromise for detection of PSC layers in lidar backscatter profiles acquired at Dumont D'Urville is a time averaging window of 1 hour typically.

2 Introduction

During winter, the polar regions do not receive sunlight and so do not benefit anymore from heating associated with the absorption of ultraviolet radiation by ozone. The infrared cooling combined with the effect of isolation provided by the polar vortex quickly generates temperatures in the polar lower stratosphere that are low enough for the formation of PSC between 15 and 30 km. PSCs play a key role in the formation of the so-called ozone hole over Antarctica at the beginning of the spring. PSCs provide reactive surfaces for heterogeneous chemical reactions that result from interactions between species in the gas phase and surfaces/volumes of PSCs solid or liquid phases. These reactions very quickly convert halogen reservoir species into ozone-destroying radicals (see for example [WMO, 2007] and [Peter, 1997]). PSCs may also play a significant role in the radiative balance of the atmosphere [Cirbus Sloan and Pollard, 1998] or [Lachlan-Cope et al., 2009]. A long term increase in PSCs might even influence the climate of the lower stratosphere. Note however that long and homogeneous observational times series of PSCs remain scarce [David et al., 2009].

Several types of PSC have been identified and are usually distinguished according to their optical properties. The optical properties depend on PSCs size distribution, state and composition that are quite variable. As the crucial parameter in the processes of formation and evaporation of PSCs is the temperature, its evolution mostly determines changes in PSC composition, phase and size distribution. PSCs can be liquid or solid, composed of nitric acid-rich mixtures or ice and have typical sizes of approximatively a micron, [Rosen et al., 1975], [Voigt et al., 2000] and [Tabazadeh et al., 1994].

A widely used remote instrument technique to detect PSCs is the lidar "LIght Detection And Ranging", ([SPARC, 2010] and [Fiocco and Smullins, 1963]). Lidar measurements consist of very short pulses of focused light, illuminating the overhead atmospheric column, with a relatively low divergence. The returning photons are collected and converted into an electrical signal. The return signal is collected and the time between the emitted laser pulses and the scattered returned signal is proportional to the altitude at which the scattering occurred. The intensity of the returned signal depends on the nature and concentration of the scatterers, [Bohren and Huffman, 1983] and [SPARC, 2010]. PSC detection is not only important for studies of the chemistry and dynamics of the polar stratosphere. It also allows to identify stratospheric profiles where only sulphuric acid aerosols particles are present ([Sing Wong et al., 2009] and [Adriani et al., 1999]) and can

be used as cloud-free reference profiles for lidar calibration ([Platt, 1979]).

The large amount of data (several thousand lidar profiles per year) makes it difficult to identify in a reliable and objective way the presence of PSC layers on every profile. Many detection methods exist in the literature ([Chang and Zhang, 2007, Gumedze et al., 2010]). Still some studies do not pay attention to stationarity properties of the signal (for example, homoscedasticity which indicates that the variance of the signal is constant) which theoretically preclude some statistical calculations of interest (see [Goldfarb and Pardoux, 2007]). Other methods rely on the use of arbitrary thresholds ([Morille et al., 2007], or [Berthier et al., 2008]), or require the a-priori knowledge of the optical properties of the scatterers, here PSC (see the work of [Chazette et al., 2001]). The present study proposes a new method to automatically detect PSC layers in a profile. The method is based on the fact that the variance of a backscatter profile is locally affected by the presence of PSC layers. PSCs are identified here in lidar profiles as an increase in the variance of the signal with an automated procedure that does not require the use of visual or ad-hoc threshold selection.

The paper is organized as follows. Section 3 briefly describes the lidar data we used. The detection procedure is explained in section 4, introducing by the way the different statistical characteristics of the lidar data. Section 5 presents and discusses the results on the application of the detection procedure to a large lidar data set. The last section is devoted to other possible applications of this detection method and concluding remarks.

3 Lidar data

The international Network for the Atmospheric Composition Changes (NDACC) is composed of worldwide remote-sensing stations monitoring the physical and chemical parameters of the atmosphere. The current study is focused on lidar data collected at the Dumont d'Urville (hereafter refered as DDU, 66°39'46"S 140°0'5"E) station in Antarctica. The lidar initially installed in 1989, provides vertical backscatter profiles of the atmosphere from several meters above the instrument to well above the top of the stratosphere, with a 5 minutes time integration. About 100-200 nights of observations are performed per year.

The retrieval process and necessary assumptions in processing lidar data from DDU are explained in details in [Chazette et al., 1995] and [David et al., 1998]. Instrumental concerns on the DDU lidar and NDACC network detection capabilities can be found for example in [Shipley et al., 1983], [Von Zahn et al., 2000] and in [David et al., 1998].

These measurements provide backscatter aerosols profiles which can contain indication of the presence of PSCs over Antarctica. The vertical resolution of the profiles is 75 meters. Since PSCs form between 15 and 30km approximately, the detection procedure is applied on the altitude range between 8 and 35km only, giving 360 data points per lidar profiles. The equation relating the received backscattered signal intensity $P(z)$ from a given z altitude, involving the extinction from the air column and particles ranging from the lidar ground level to the backscattering z altitude is given by

$$P(z) = F_0 \beta(z) \frac{K}{z^2} \exp\left[-2 \int_{z_0}^{z} \alpha(z')dz'\right], \tag{3.4}$$

where $P(z)$ is typically the lidar power incident on receiver from z (typically a flux photons : number of photons per unit time and unit surface), F_0 is the laser pulse energy, $\beta(z)$ is the total aerosol and molecular backscatter coefficient, Kencompasses the various instrumental constants (including area of the lidar receiver) and $\alpha(z)$ is the total extinction coefficient (molecules + particles). In particular, the presence of clouds layers modify the scattering and extinction properties along the optical path of the laser beam. The resolution of this equation is widely discussed in literature (see for example [David, 1995], [Collis and Russell, 1976], [Fierli et al., 2001] and [David et al., 2004]). This gives rise to both theoretical and instrumental issues. [Fernald et al., 1972] and [Klett, 1981] and [Klett, 1985] identified a first order Bernouilli differential equation and stated on the formalism of its solution. The critical assumption is the a-priori knowledge of the ratio between extinction and backscattering, the so-called lidar ratio. The values of this ratio depend on the particle type, being either aerosols, cirruses, or PSCs. With known lidar ratios, an objectivity issue still remains in the selection of the altitudes separating the different particle types along any lidar profile. This step has to use quantifiable and objective criteria to ensure the reliability of lidar time series. This is the substance of the present paper.

4 An procedure to detect PSCs

An example of a cloud-free profile is displayed in the top left hand corner of Figure 3.2, this profile was measured on 2008/04/17 over the DDU station. Typically, the backscattered signal decreases sharply with the increasing altitude between 8 and 35km. Every backscatter profile exhibits an interesting statistical feature : the variance (calculated from the difference between the raw and smoothed profiles) is never constant, and

varies with altitude (see panel b of Figure 3.2). A signal with varying mean and/or variance is called a heteroscedastic signal. Most of the cloud-free (i.e. background) variance originates from instrumental noise and, possibly, some natural short-term variability of the atmosphere.
The presence of a PSC layer in a profile (panel d of Figure 3.2, profile measured on 2008/08/23) generates a local increase in the variance, as illustrated in the panel 3.2-e which shows the same profile as in 3.2-b after removing the smoothed profile (i.e. the low frequency component of the signal ; thereafter referred as smoothed signal or trend). The lower altitude of 8km was chosen to prevent including high-altitude cirrus clouds in the variance estimation.

Our procedure detection is based on these three characteristics (i.e. the trend, the decreasing variance and the transient variance break) and requires three steps in the signal processing. The first step is the stationarization of the signal. That means removing the trend and controlling the variance. This first operation allows us to proceed, in the second step, to the maximum likelihood estimation of the parameters of the model, and then estimate the more likely altitude range of a PSC layer. The last step uses a Fisher-Snédécor test to rule if the detection of PSC is statistically significant.

Based on the characteristics of the lidar backscatter profiles described previously, the raw signal P_{raw} is modelled with a linear combination of signals including random Gaussian variables

$$P_{raw} = P_{trend} + P_{cloud} + P_{back} \tag{3.5}$$

where P_{trend} describes the trend of the signal (low frequency component of the signal). P_{cloud} describes the signal fluctuations generated by the PSC ; this PSC signal is null except between two boundaries, the top and bottom altitudes of the PSC layer, where it is modelled with a zero-mean Gaussian variable whose distribution is usually denoted by $\mathcal{N}(0, \sigma^{*2}_{cloud})$ with 0 being the mean and σ^{*2} being the variance. Finally P_{back} describes the heteroscedastic (i.e. variance is not constant) background signal which is modelled with a zero-mean Gaussian variable whose distribution is denoted by $\mathcal{N}(0, \sigma^{*2}_{back})$; σ_{back} is the altitude-dependent background variance which is found to decrease approximately linearly with increasing altitude (Figure 3.2-b).

4.1 Stationnarization procedure

As explained above, a backscatter profile is obviously not stationary, neither in mean nor in variance. The smoothing of the signal P_{trend} is carried out using a centred moving average filter of vertical length p with p being the number of points of averaging window. Once the trend is estimated, it is subtracted from the raw signal to generate a zero-mean signal P^{hf} given by

$$P_{hf} = P_{raw} - P_{trend} = P_{cloud} + P_{back}. \tag{3.6}$$

The residuals P_{hf} are the high-frequency component of the signal. They are heteroscedastic and so P_{hf} is non- stationary. However, analysis of P_{hf} in a large number of backscatter profiles show that the standard deviation of the backscatter profile, denoted by σ_{back}, varies regularly with altitude and that its altitude dependency can be accurately reproduced by a linear trend over the cloud-free altitude ranges ; this allows us to remove the altitude dependency of the variance in P_{hf} in order to generate a stationary signal. The standard deviation σ_{back} of P_{hf} at a cloud-free altitude z is given by

$$\sigma_{back} = a + bz. \tag{3.7}$$

It is worth pointing out that, over the cloud altitude range, the total variance is expected to be higher because it will be the sum of the background variance σ_{back} and of the cloud variance σ_{cloud}. After estimating the constants a and b using a common least square fitting approach, the final step to stationarize the signal is to divide P_{hf} by its own variance σ^2_{back}. This step is similar to an altitude-dependant normalisation and can be expressed as

$$P^* = \frac{P_{hf}}{\sigma^2_{back}}. \tag{3.8}$$

P^* is homoscedasctic and has units of $power^{-1}$ whereas P_{raw} has units of $power$. The exponent * is always used here to refer to quantities derived from the stationarized signal P^* (generated by the altitude-dependent normalisation given by Equation (3.8)). Once the signal is stationarized, the resulting distributions of P^* can be considered as independent and identically distributed, and it remains constant over the cloud-free altitude ranges (see panel c of Figure 3.2).

The analysis of a large number of backscatter profiles indicates that the distribution of the stationarized signal P^* can be assumed to be Gaussian (zero-mean and constant variance) over two distinct altitude ranges, outside and inside the PSC layer. Outside the PSC layer, the distribution is denoted by $\mathcal{N}(0, \sigma^{*2}_{back})$. The signal P^* displays a higher variability within a PSC layer (see Figure 3.2-f) and the distribution of P^* within a PSC layer is denoted by $\mathcal{N}(0, \sigma^{*2}_{cloud})$. When analysing the results, it must be kept in mind that σ^2_{cloud} refers to the variance of P_{hf}, the high-frequency component of the backscatter profile, whereas σ^{*2}_{cloud} refers to the variance of P^*, the stationarized P_{hf}.
The entire previous procedure is illustrated in Figure 3.2 for a cloud-free profile measured on 2008/07/08 and for a profile where a PSC layer appears between 18 and $21,5$ km on 2008/07/09. The three panels on the top of Figure 3.2 correspond to the cloud-free profile monitored on 2008/07/08 : the panels 3.2-a and 3.2-b show the raw profile P_{raw} and the variance of P_{hf} (=raw profile - smoothed profile) respectively. Panel 3.2-c shows the stationarized profile P^* resulting from the three-step processing described above. The profile P^* appears as a somewhat constantly distributed signal over the cloud-free altitude ranges, while, in the case of a PSC layer (the three bottom panels), the variance sharply increases between the two cloud boundaries that have to be estimated.

4.2 PSC parameters estimation by likelihood maximisation

This section explains the likelihood maximisation procedure on the signal P^* in order to determine the most likely altitude range of a possible PSC layer. The M_0-model assumes the profile does not contain a PSC. Conversely, the alternative M_1-model assumes there is a PSC somewhere in the profile between two altitudes τ_b and τ_t, to be estimated representing respectively the bottom and top altitude of the PSC layer.
Thanks to the stationarisation procedure, the signal P^* can be assumed to be an independent and identically distributed (iid) Gaussian with a higher variance within the PSC layer. The two models are presented as

$$M_0 : P^* \text{ variance denoted by } \sigma^{*2}_{out} \text{ does not vary with altitude,} \tag{3.9}$$

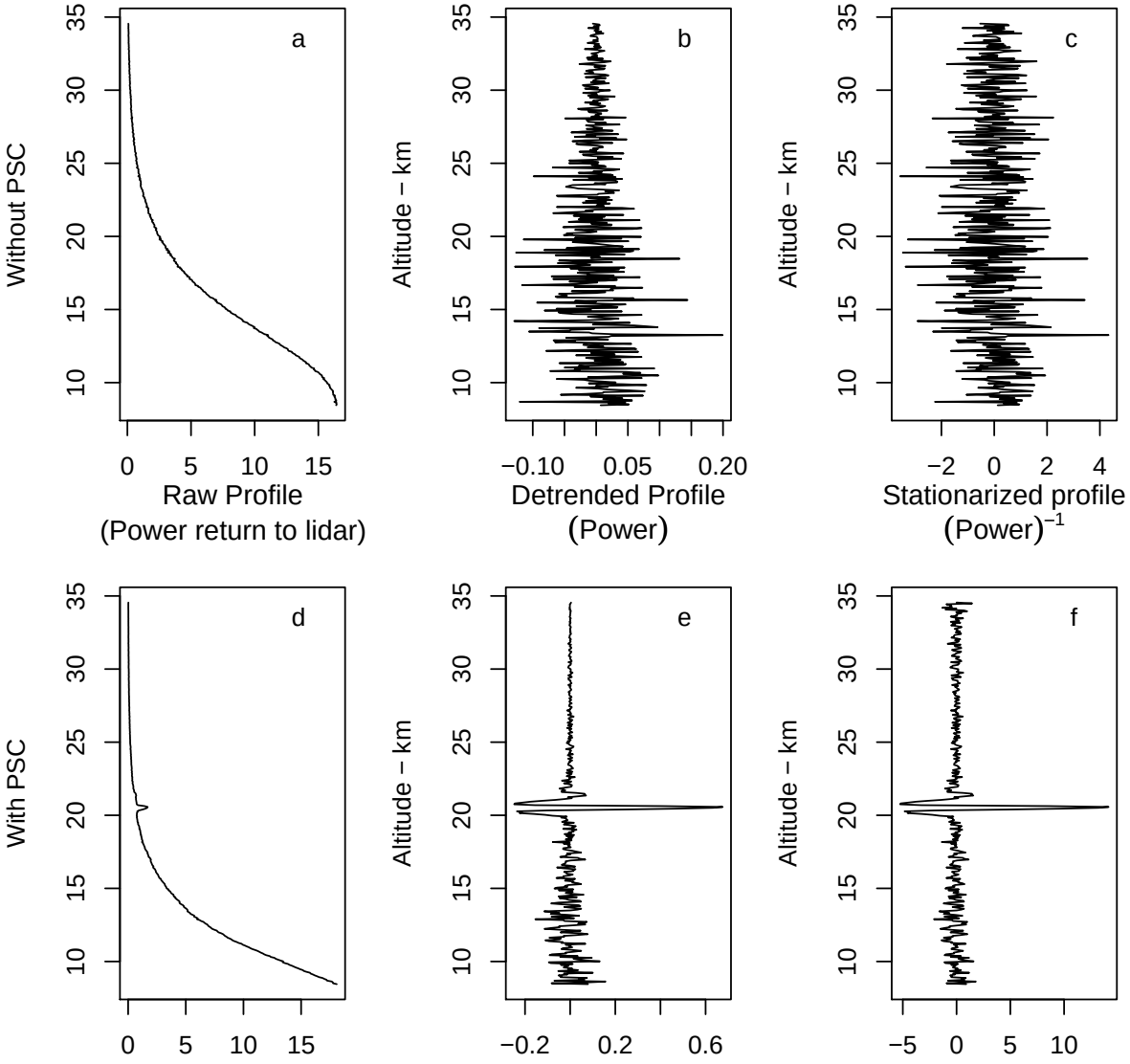

FIGURE 3.2 – Our stationarisation procedure. The three plots on the top correspond to the different steps of stationarisation for a clear sky profile monitored on 2008/04/17, while the three plots on the bottom illustrate the procedure for a profile monitored on 2008/08/23 and displaying a PSC between 16km and 24km. Note that the scales of the panels are different.

whereas

$$M_1 : P^* \text{ variance equals to } \sigma^{*2}_{in} \text{ within the altitude range } [\tau_b, \tau_t[\text{ and } \sigma^{*2}_{out} \text{ otherwise,} \tag{3.10}$$

with the index out referring to the domain *outside* the PSC layer and in referring to the domain *inside* the PSC layer. As M_0 is included in M_1 (by considering $\sigma^{*2}_{in} = \sigma^{*2}_{out}$), all models corresponding to M_0 also correspond to M_1. In this case the two altitudes τ_b and τ_t still exist but do not have any influence on signal P^*.

The underlying likelihood of model M_1 following (3.10) is given by

$$\begin{aligned}\mathcal{L}(P^*; \sigma^*_{out}, \sigma^*_{in}, \tau_b, \tau_t) = \\ -n \log(\sqrt{2\pi}\sigma^*_{out}) + (\tau_t - \tau_b) \log \frac{\sigma^*_{out}}{\sigma^*_{in}} - \frac{1}{2}\left[\sum_{z \notin [\tau_b, \tau_t[} \frac{[P^*(z)]^2}{\sigma^{*2}_{out}} + \sum_{z \in [\tau_b, \tau_t[} \frac{[P^*(z)]^2}{\sigma^{*2}_{in}} \right],\end{aligned} \tag{3.11}$$

where σ^*_{out}, σ^*_{in}, τ_b and τ_t are the parameters that need to be estimated, and n is the number of altitude range.

The details of the calculation giving (3.11) are given in Appendix 7.1. The procedure aims to estimate the previous parameters by maximising the likelihood (3.11). This maximisation of (3.11) has to be done under the obvious constraints that the bottom altitude of the PSC layer has to be lower than the top altitude and that these two altitudes have to be found within certain boundaries (i.e. the bottom altitude is above 15km and the top altitude is below 30km). The final constraint is that the variance of the signal within the cloud layer (σ^*_{in}) has to be higher or equal to the variance of the cloud-free domain (σ^*_{out}), or, more precisely, that the two variances have to be equal when there is no PSC. Overall the maximisation under constraints can be expressed by

$$\begin{aligned}&\underset{\sigma^*_{out}, \sigma^*_{in}, \tau_b, \tau_t}{\arg\max} \ \mathcal{L}(P^*; \sigma^*_{out}, \sigma^*_{in}, \tau_b, \tau_t) \\ &(a) \ \ 0 \leq \sigma^*_{out} \leq \sigma^*_{in} \\ &(b) \ \ 15\text{km} \leq \tau_b \leq \tau_t \leq 30\text{km}.\end{aligned} \tag{3.12}$$

There are a number of difficulties in solving (3.12) (likelihood $\mathcal{L}$ not continuous with

respect to τ_b and τ_t (see 3.11), taking into account the constraints, the number of parameters,à). However, the resolution can be simplified. Instead of having the 4 parameters (σ^*_{out}, σ^*_{in}, τ_b and τ_t) as control variables in this maximisation problem with constraints, $\mathcal{L}$ is only maximised with respect to τ_b and τ_t using as σ^*_{out} and σ^*_{in} as fixed parameters that have been estimated previously. Then, once $\mathcal{L}$ is maximised, the corresponding values of τ_b and τ_t are used to recalculate σ^*_{out} and σ^*_{in} which are in turn used in a new resolution of (3.12). At the end of each iteration, the values of τ_b and τ_t estimated by the resolution of (3.12) are compared to the values of τ_b and τ_t estimated in the previous iteration and used to calculate σ^*_{out} and σ^*_{in} (inputs to the resolution of (3.12)). As long as the input and estimated values of τ_b and τ_t are significantly different, this procedure is repeated. It is found to converge after fewer than 5 iterations in most cases.

The estimation of the variances is performed using the definition of the empirical variance (see [Sprinthall, 2009]) by splitting the signal in two intervals. The first interval corresponds to the cloud-free domain $[z_1, \tau_b[\cup[\tau_t, z_n]$. The second one corresponds to the PSC domain $[\tau_b, \tau_t[$. The respective variances of these intervals (i.e. inside and outside) are given by

$$\begin{aligned} \hat{\sigma}^{*2}_{out} &= \frac{1}{n-(\tau_t-\tau_b)} \sum_{z\in[z_1,\tau_b[\cup[\tau_t,z_n]} [P^*(z)]^2, \\ \hat{\sigma}^{*2}_{in} &= \frac{1}{(\tau_t-\tau_b)} \sum_{z\in[\tau_b,\tau_t[} [P^*(z)]^2. \end{aligned} \tag{3.13}$$

where τ_t and τ_b are expressed in units of number of datapoints in the vertical profile instead of km with 8 km being the origin. These two estimates correspond to the values of σ^*_{out} and σ^*_{in} which maximize equation (3.11), when concidering τ_t and τ_b as constant.

The first estimates $\hat{\sigma}^*_{out}$ and $\hat{\sigma}^*_{in}$ (used as inputs in the first resolution of (3.12)) are calculated assuming that the cloud-free altitude ranges cover below 15km and above 30 km because PSCs are usually not observed at those altitudes. This choice of altitude ranges is rather arbitrary. Nonetheless, it has no influence on the final estimation because the iteration procedure recalculates recursively the cloud and cloud-free altitude ranges. After a few iterations, the estimates of $\hat{\sigma}^{*2}_{out}$, $\hat{\sigma}^{*2}_{in}$, $\hat{\tau}_b$ and $\hat{\tau}_t$ do not change anymore. Further investigations on the robustness of the estimation are discussed in part 4.4.

As the cloud altitude range corresponds to discrete values (vertical resolution of 75

m), the maximisation of $\mathcal{L}$ with respect to τ_b and τ_t be computed numerically. It is not necessary to calculate the entire $n \times n$ matrix, with n being the total number of discrete altitudes. First, the constraint (3.12-b) $\tau_b \leq \tau_t$ means that only half the calculation of the matrix is needed. Second, the fact that form between 15km and 30km further limits the calculations to $\tau_b > 15$km and $\tau_t < 30$km. An example of matrix ($\mathcal{L}$ as a function of τ_b and τ_t) is provided in Figure 3.4.

4.3 Statistical significance of the parameters estimation by a transient shift test

Once convergence is achieved, the maximum likelihood algorithm provides the best estimates of the parameters (cloud altitude range and variances over the cloud and cloud-free domains), assuming there is a PSC layer. However, it does not check the likelihood of the existence of the PSC layer. Now it is time to test the statistical significance of the PSC detection as defined by these parameters : ($\hat{\tau_b}$ and $\hat{\tau_t}$) representing the best estimates of the bottom and top altitudes of a hypothetic PSC and $\hat{\sigma}^{*2}_{out}$ and $\hat{\sigma}^{*2}_{in}$ representing the best estimates of the variances in the interval $[z_1, \tau_b[\cup[\tau_t, z_n]$ and in the interval $[\tau_b, \tau_t[$ respectively. A test is needed to rule whether the detection of a PSC layer is statistically significant.

The two-hypothesis model can be reduced to the problem to know whether $\hat{\sigma}^{*2}_{out} = \hat{\sigma}^{*2}_{in}$ or $\hat{\sigma}^{*2}_{in} > \hat{\sigma}^{*2}_{out}$, or similarly to know if, statistically, the variability inside and outside the PSC can be considered as equal or if the variability is statistically significantly higher in the *inside* interval than the one in the *outside* interval. This last case would indicate the presence of a PSC.

A fisher-Snédécor test handles this problem by considering the ratio of the squared variances of each samples (see [Mood, 1974]). The ratio allows to test the equality of the variance of two independent samples. Two samples are created from the values of P^* split in the two different intervals with the test taking into account the different sizes of the two samples. The ratio is then given by

$$F_{n_1-1,n_2-1} = \frac{\hat{\sigma}^{*2}_{in}}{\hat{\sigma}^{*2}_{out}}, \tag{3.14}$$

where $\hat{\sigma}^{*2}_{in}$ and $\hat{\sigma}^{*2}_{out}$ both follow a $\chi^2_{n_i-1}$ distribution, as weighted sums of squared Gaussian variables, (with n_1 being the sample size of the *inside* interval and n_2 the sample size of

the *outside* interval).

This implies that F follows a Fisher distribution with (n_1-1, n_2-1) degree of freedom. As commonly done in statistics, the decision is made using a fixed confidence rate of 97%. This test ultimately decides on the existence of a PSC layer.

4.4 Estimation of bias and detection limit using simulated data

The purpose of this section is to evaluate the performances of the detection algorithm on perfectly characterized data that are generated numerically. In such a configuration, one can assess the ability of the algorithm to detect and quantify a-priori known signals in the profiles. The characteristics are chosen such that they mimic typical characteristics of lidar profiles. The aims of this type of numerical experiment are, for instance, to identify possible biases and estimate a detection limit of PSCs.

Non-stationary signals are first simulated numerically. Signals representative of average background backscatter profiles are generated by combining a smoothed profile average backscatter profile and a heteroscedastic (i.e. altitude-dependent) Gaussian noise (=$\mathcal{N}(0, \sigma^2_{back})$) ; $\sigma_{back} = 3 - 2z/360)$, for $z \in [1, 360]$ with z expressed in units of number of points in the vertical profile (8 km corresponding to the origin). Then, between two altitudes, corresponding to the bottom and the top altitudes of a PSC layer, another Gaussian noise with a greater variance (=$\mathcal{N}(0, \sigma_{cloud})$) is added to the background profiles. An example of profile simulated by adding a cloud variance σ_{cloud}=20 between $19,7$ and $22,3$ km is shown in Figure 3.3. The detection algorithm is applied to this simulated lidar profile ; Figure 3.4 shows the likelihood (see Equation (3.11)) as a function of the cloud altitudes. The best estimation of the cloud altitudes is provided by the maximum of the likehood, indicated by the open circle on Figure 3.4 and by the dotted lines in Figure 3.3. The retrieved cloud bottom altitude is underestimated by about 300 m (corresponding to 4 data points for the 75m vertical resolution of the profiles) and the cloud top altitude is overestimated by the same amount.

The performances of the algorithm are then tested for a wide range of cloud variance values in order to characterise further biases and estimate the detection limit which is expected to depend both on the cloud-to-background variance ratio and on the length of the moving average window, p (used to smooth the raw lidar backscatter profiles (see 3.2)). Note that, for each value of cloud variance σ_{cloud} considered, 500 profiles are simulated and treated by the detection algorithm.

Figure 3.5 shows the PSC altitude range, $\hat{\tau}_b$ and $\hat{\tau}_t$, estimated by the detection algorithm as a function of the cloud variance σ_{cloud} which is added to the simulated background profiles between $19,9$ and $23,5$ km. The profiles are smoothed with a moving average window of length $p = 10$. The size of the boxes (bounds indicating 25th and 75th percentiles), what draws an overview of the distribution pattern, indicates that half the estimates are concentrated in a 200meters-wide interval typically. There are two distinct regions in Figure 3.5. For σ_{cloud} smaller than σ_{back}, the retrieved values of the PSC altitude range vary substantially with many outliers at very small values of σ_{cloud}. This suggests that the estimation of the cloud altitude range is not fully reliable when $\sigma_{cloud} < \sigma_{back}$. In contrast, for σ_{cloud} greater than σ_{back}, $\hat{\tau}_b$ and $\hat{\tau}_t$ vary little and there are not a single outlier. The same features and evolution are found at the top and bottom cloud altitude. The estimation appears to be robust. However, the retrieved values exhibit a bias of about 300 m with respect to the cloud altitude range where the variance was enhanced compared to the background variance. The bias is positive at the top cloud altitude and negative at the bottom.

This bias in the estimated cloud altitudes is caused by the way the profiles are smoothed. Let's recall that a PSC is generated by enhancing the variance on a simulated background profile within a given cloud altitude range. As the smoothed raw profile (i.e. trend P_{trend}) is estimated with a moving average, the smoothed raw profile differs from the smoothed background profile, not only within the cloud altitude range (from τ_b to τ_t), but also in the vicinity of the cloud boundaries. Indeed, the moving average being of length p, the trend P_{trend} is expected to be modified over an altitude range exceeding the cloud altitude range by about 375 m ($75m \times p/2$, where 75m is the vertical resolution) on each side of the cloud boundaries. As a result, the high-frequency component P_{hf} (=$P_{raw} - P_{trend}$) and the associated variance are artificially enhanced by the presence of a PSC layer from $\tau_b - p/2$ altitude to $\tau_t + p/2$) altitude. As the PSC detection algorithm is based on the detection of changes in the variance, the estimated cloud bottom (top) altitude is found to be lower (higher) than in the simulated raw backscatter profile. Figure 3.5 illustrates quite well this small bias of the detection algorithm. It means that, for an accurate determination of the cloud altitude range, the bias has to be removed from the cloud altitude range estimated by the algorithm. It is also necessary for the cloud variance σ_{cloud} to be at least of the order of the background variance σ_{back} in order for the algorithm to detect and reliably estimate the cloud altitude range. The level of the background variance in the profile can also be interpreted as the detection limit of the algorithm.

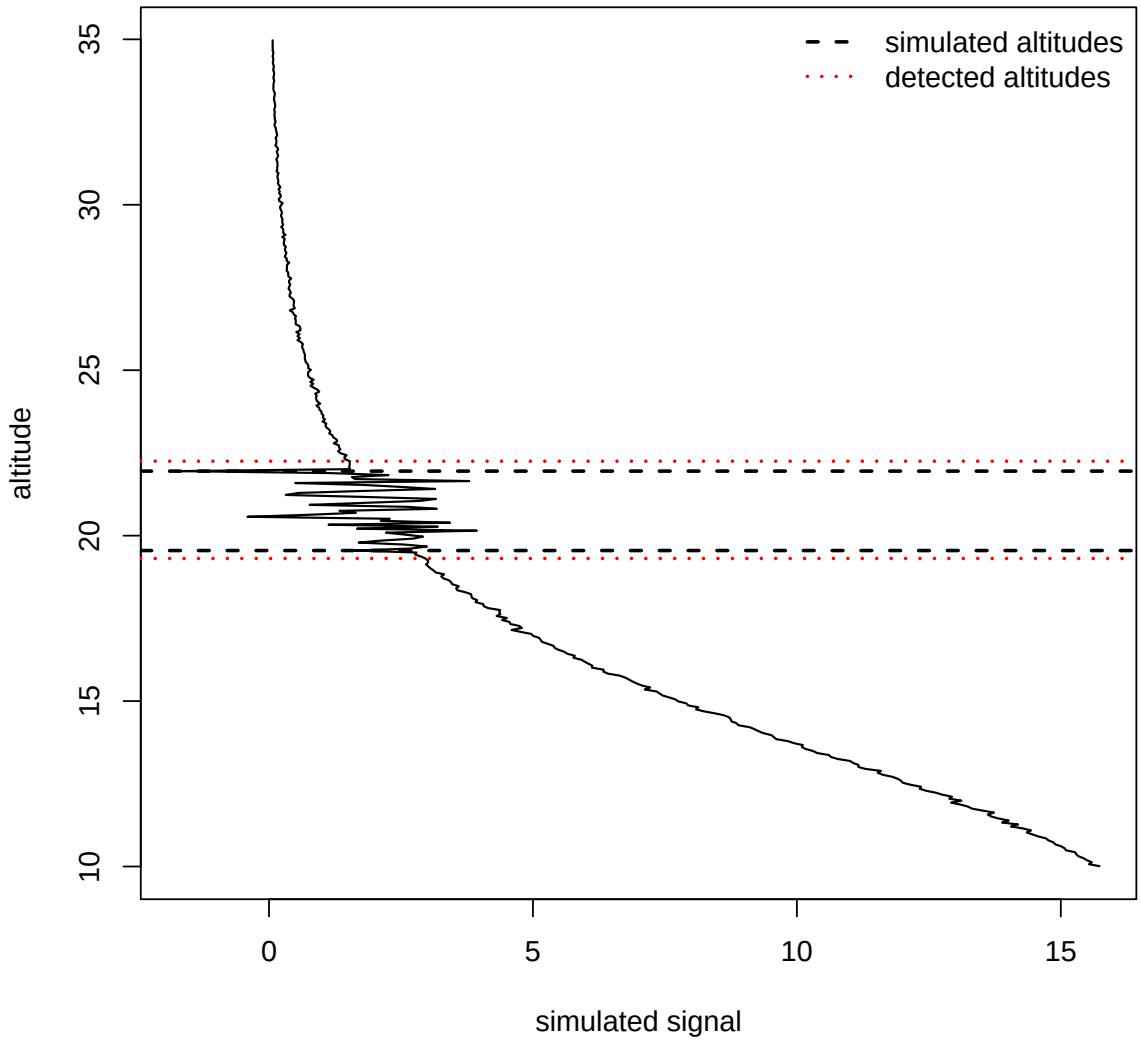

FIGURE 3.3 – Detection of a PSC in a simulated backscatter profile (black line). The cloud bottom $\hat{\tau}_b$ and top $\hat{\tau}_t$ altitude estimated by the detection algorithm are indicated with the dotted lines ; the actual cloud altitude range, as simulated in the profile, are indicated with the black dashed lines.

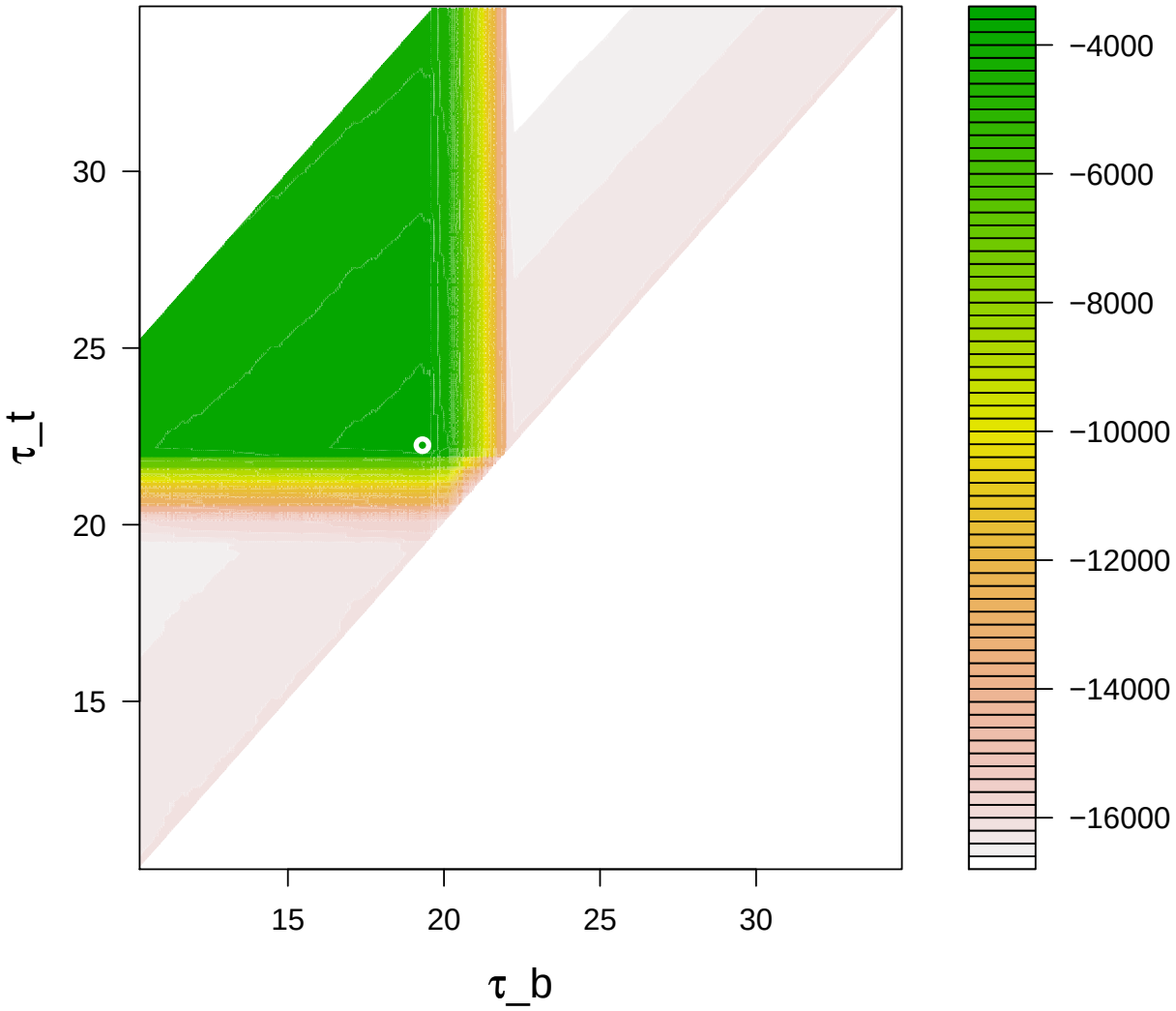

FIGURE 3.4 – The likelihood $\mathcal{L}$ as a function of the cloud bottom τ_b and top τ_t altitude for the simulated profile of Figure 3.3. The maximum of $\mathcal{L}$ is indicated with an open circle.

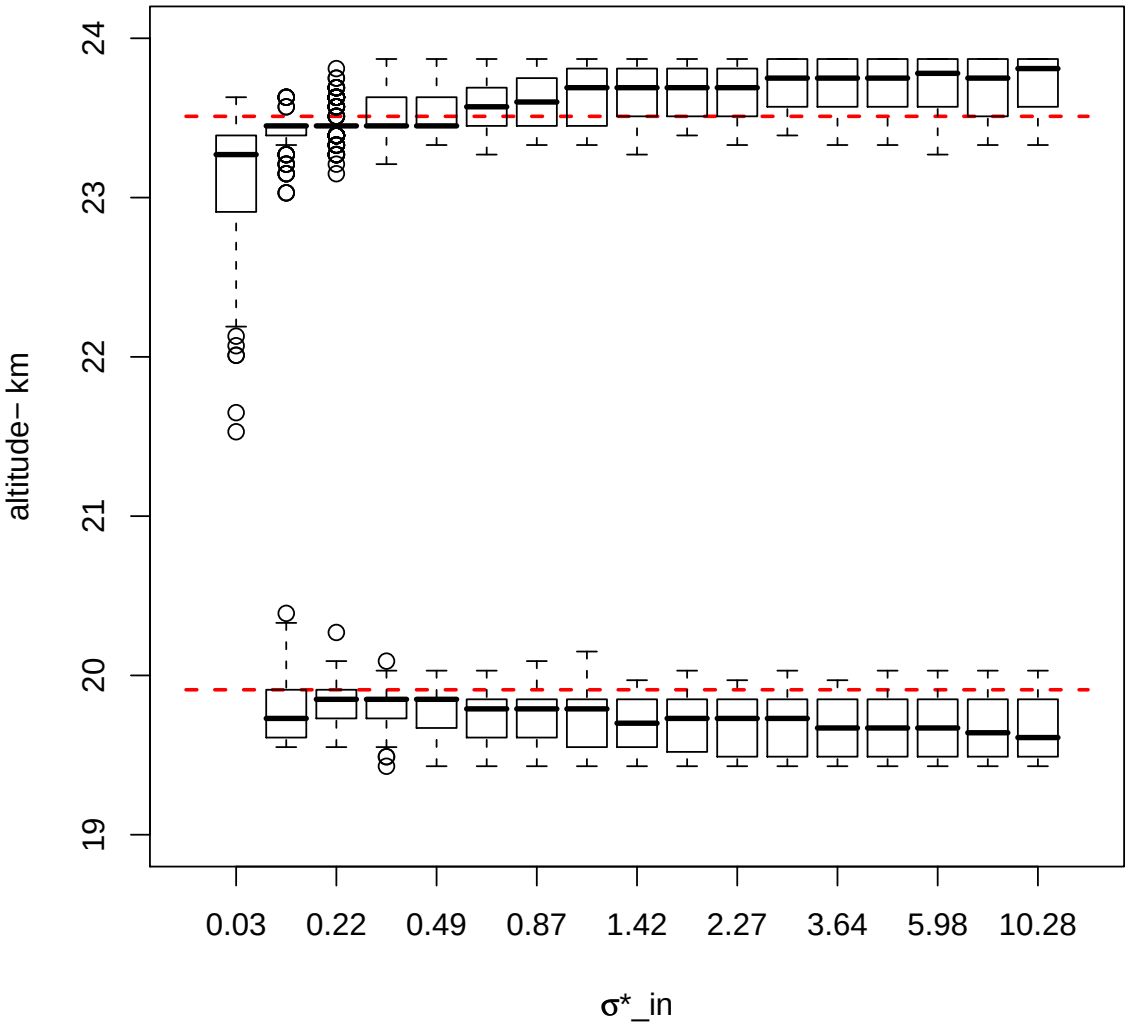

FIGURE 3.5 – Boxplot of the PSC altitude range, $\hat{\tau}_b$ and $\hat{\tau}_t$, estimated by the detection algorithm as a function of the cloud variance σ_{cloud} which was added between $19,9$ and $23,5$ km to the simulated background profiles. The median value (thick horizontal black bar), 25th and 75th percentiles (lower and upper box bounds respectively), and the lowest and highest data within 1,5 interquartile range of the lower and upper quartile respectively (lower and upper whiskers respectively) are also indicated. The outliers (i.e. data not included between the whiskers) are plotted as open circles. The actual PSC altitude range is indicated with two dashed horizontal lines ($19,9$ and $23,5$ km).

5 The effect of temporal averaging of profiles using real data.

This section describes the study of real backscatter profiles measured at the DDU station. As a first example, the detection of a PSC over DDU on July 9th 2008 is presented in Figure 3.6. The estimated cloud altitude range (between 18.1km and 21.15km) is indicated with the dashed lines. For the same example, the evolution of the likelihood $\mathcal{L}(P^*; \sigma^*_{out}, \sigma^*_{in}, \tau_b, \tau_t)$ is plotted as a function of the cloud bottom τ_b and top τ_t) altitude in Figure 3.7. The maximum of $\mathcal{L}$ is represented with an open circle and indicates the best estimates of the PSC bottom and top altitude. Overall, the processing of measured backscatter profiles by the algorithm gives results that are very similar to those obtained with simulated profiles (see Figure 3.4). The statistical significance of these estimates is calculated using the Fisher Snedecor test of Equation (3.14) with the 97% confidence rate.

The detection algorithm is applied to lidar aerosol backscatter profiles measured between March and October 2008. Lidar aerosol profiles are available at a 5 minutes resolution corresponding to the measurement time integration. The total number of profiles is 3857. In the literature, before analysis, backscatter profiles are usually averaged over several hours. The averaging allows to minimise the measurement noise and, therefore, make it easier to detect the aerosol/cloud signals. In essence, it is a way of reducing the background variance and hence improving detection. However, the averaging process also has negative consequences. It degrades the temporal resolution. And, it can reduce the cloud signal/variance when the cloud characteristics are not stable over the averaging window. That is the case for rapidly varying PSC events. The averaging can lead to profiles with radically different characteristics (different PSC variance and altitude ranges, profiles absence of PSCs) being averaged together. The length of the averaging window represents a compromise between the benefit of minimising the instrumental noise and the detrimental effects of degrading the temporal resolution and attenuating the cloud signal.

The consequences of averaging the profiles is illustrated in Figure 3.8 where the altitude range of PSC layers detected by the algorithm between June and September 2008 are reported. Each panel corresponds to PSC detections carried out over different averaging intervals : 10 mn, 30 mn, 1 hr, 2 hr, 4 hr, 6 hr, 12 hr and 24 hr. All the detection results are compared with the 5 mn interval detections (the first top panel) that are indicated in grey on every other panels. The dots at the bottom of each panel indicate the average profiles processed by the algorithm. The larger the averaging interval is, the smaller the number of data (average profiles) is, the sparser the dots are. The results for March, April, May

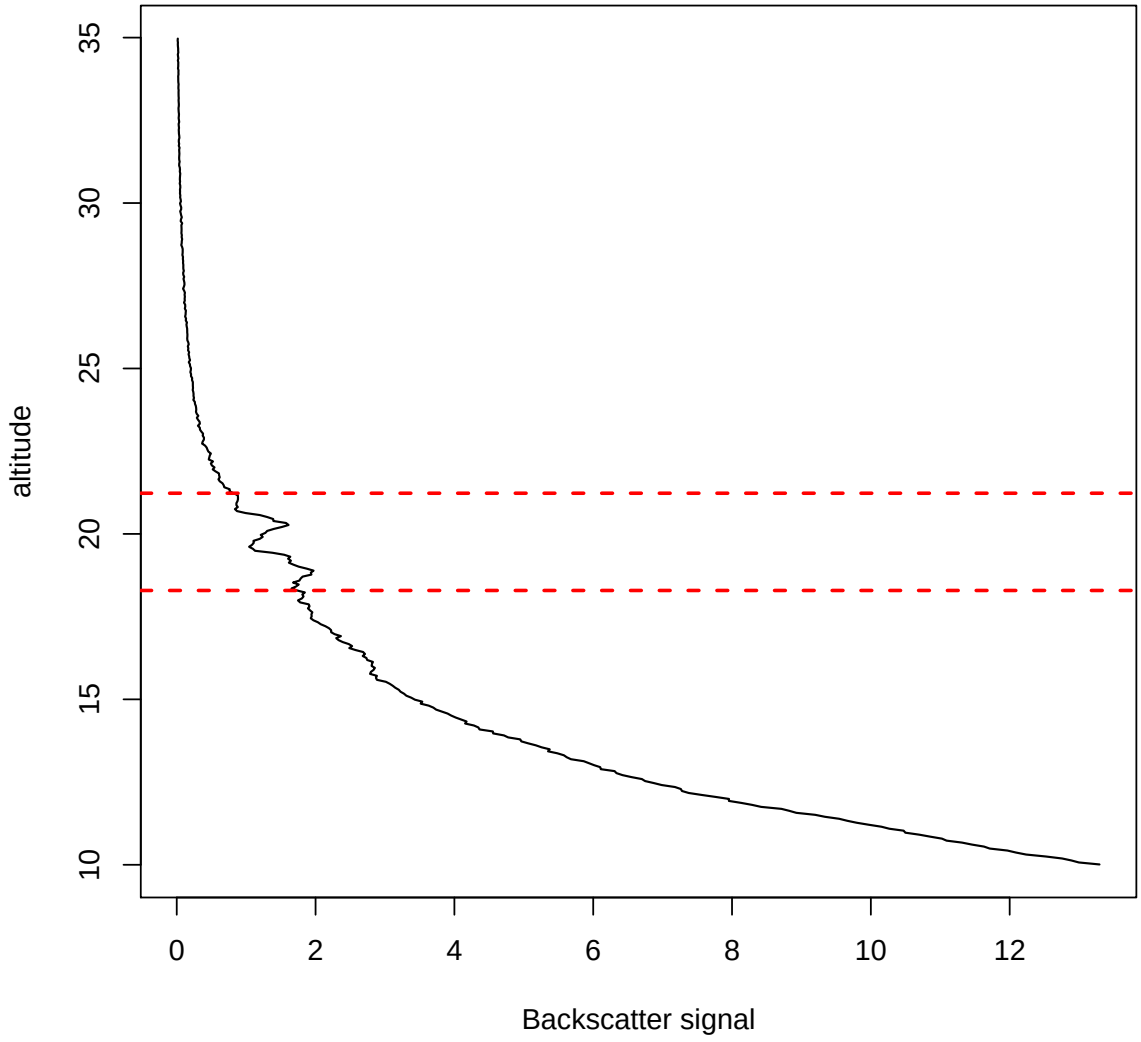

FIGURE 3.6 – Detection of a PSC between and in a 2008/07/09 profile (black line). The estimated cloud bottom altitude ($18.1km$) and top altitude ($21.15km$) are indicated with the dashed lines.

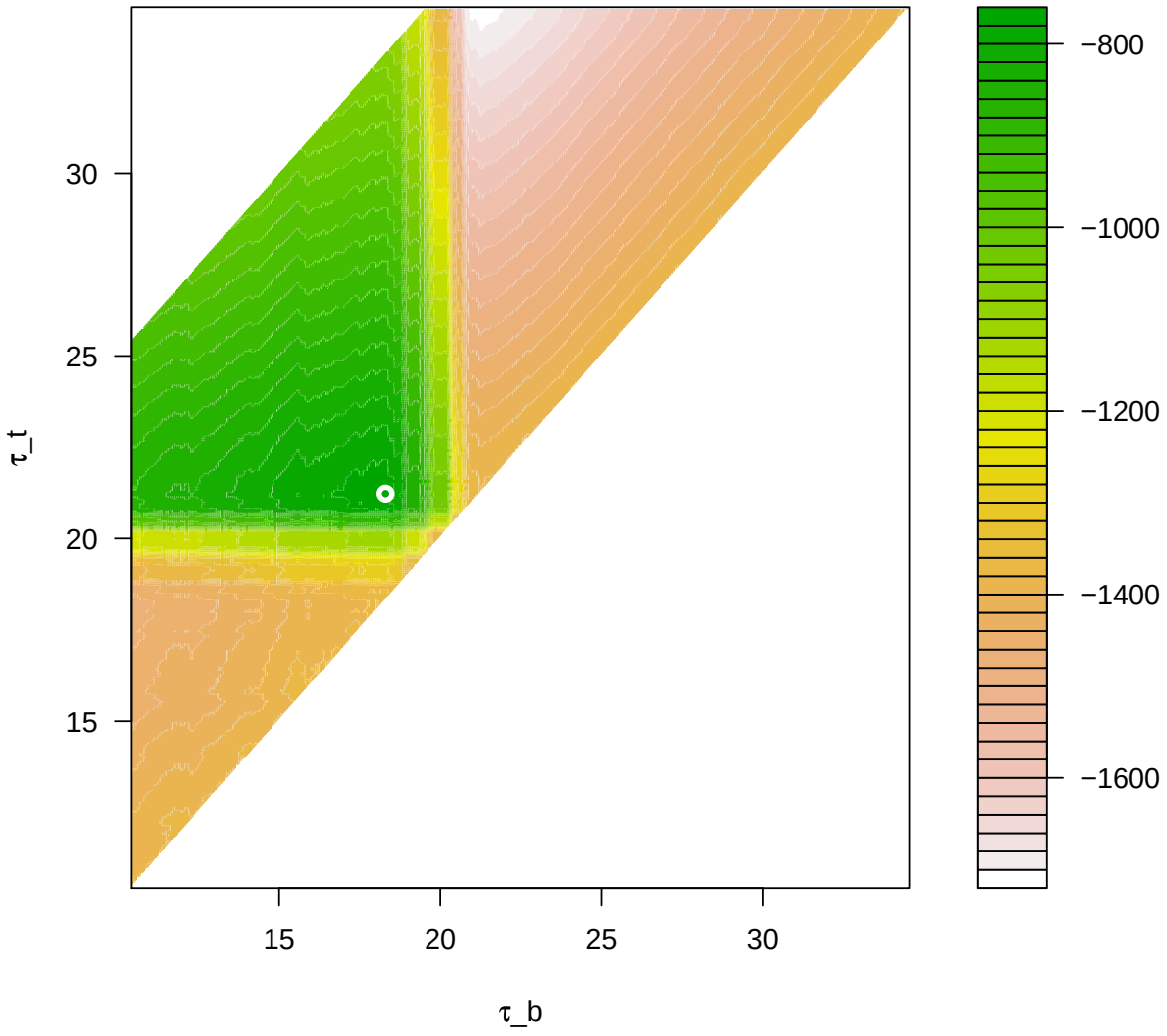

FIGURE 3.7 – The likelihood $\mathcal{L}$ as a function of the cloud bottom τ_b and top τ_t altitude for the measured backscatter profile of Figure 3.6. The maximum of $\mathcal{L}$ is indicated with an open circle.

and October 2008 are not shown because no PSCs were detected during those months except once, in May, on a 10 mn average. This detection is clearly a false positive because PSCs do not form above DDU during this period and no PSC was detected at 5 and 30 mn averaging intervals. The fluctuations from the background noise can very exceptionally (1 out of 1228) generate false positive detection at very short intervals.

The global temporal pattern of detections remains similar from a panel to another. The number of PSC detections decreases when the lidar averaging interval increases. It is expected because, at the same time, the temporal resolution and the number of profiles decrease. Note, however, that the decrease in the number of detections is stronger than expected. In addition, there is a tendency to detect thinner PSC layers when longer averaging intervals are considered. These effects start to be most significant when the averaging interval exceeds 2 hrs. For the longest averaging intervals (6 hr and beyond), some PSC layers seen on short averaging intervals are not detected anymore. It is due to the fact that, over some periods, the PSC signals are so attenuated by the averaging of mixed profiles that the algorithm is not able to detect them anymore. The effect of averaging on the signal variance can be analysed in a more formal way with the following relationship which gives the total variance of the average of two signals,

$$Var(\frac{1}{2}(P_1 + P_2)) = \frac{1}{4}Var(P_1) + \frac{1}{4}Var(P_2) + \frac{1}{2}Cov(P_1, P_2), \tag{3.15}$$

where P_1 and P_2 are two profiles.

Let's consider separately the calculation inside and outside the PSC layer. Outside the PSC layer, the covariance term (i.e. $Cov(P_1, P_2)$) should be rather constant and small compared to the first 2 terms because the background variance mostly originates from instrumental noise that is characterised by a weak temporal correlation. On the other hand, inside the PSC layer, the PSC signal is expected to exhibit longer and stronger temporal correlation whose timescales are given by the persistence of PSC events seen over DDU ; in other words, how long a PSC event typically lasts over DDU. When the profiles to average are separated by a time interval shorter than the PSC correlation timescales (and so PSC profiles with similar characteristics are averaged), the positive correlation between the profiles inside the PSC layer ensures that the inside variance decreases less quickly than the outside variance with averaging. Since the detection relies on the ratio between the inside and the outside variance, the averaging has a positive effect on the detection. For

example, there is a wide PSC layer clearly detected mid-May at short averaging intervals. However, this layer is very thin, barely detected, at the original 5 mn interval, indicating that the background noise was too strong to detect the PSC signal in the original profiles but that the averaging initially reduces the noise more than the PSC signal to make it detectable. At the largest averaging intervals, this PSC layer is not detected.

When the profiles to average are separated by a time interval beyond the PSC correlation timescales (and so profiles with completely different characteristics are averaged), the positive correlation disappears on average and the covariance ($Cov(P_1, P_2)$) inside the PSC layer should decrease with increasing averaging time intervals (then so does the variance $Var(\frac{1}{2}(P_1 + P_2))$). As a result, PSC signals become more difficult to detect in the background noise for large averaging time intervals. This attenuation effect of the averaging starts to be noticeable just on the inner edges of PSC layers where the variance is not very much higher than the outside variance. This explains why the detected PSC layers become thinner when the averaging interval is increased. For long time intervals, 6 hrs and beyond, the PSC variance can become so weak over entire PSC layers that they are completely missed by the algorithm. According to Figure 3.8, the most reliable and robust results for 2008 are obtained between 30 and 2 hrs intervals. Overall, the best compromise between the temporal resolution and the accuracy of the detection seems to be an averaging interval of 1 hr typically.

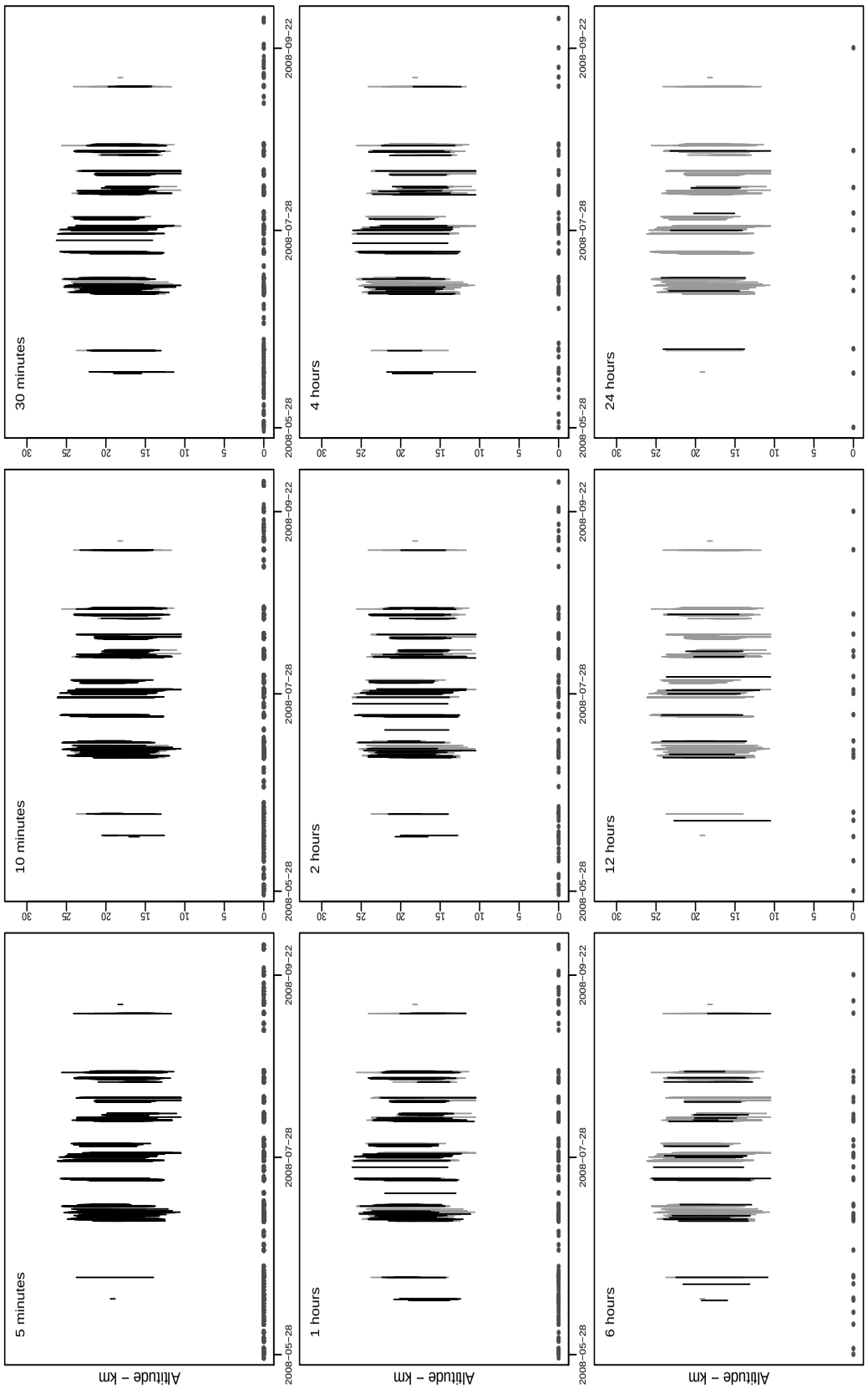

FIGURE 3.8 – Altitude range of PSC layers detected as a function of time, between June and September 2008. Each panel corresponds to PSC detections carried out over different averaging intervals : 10 mn, 30 mn, 1 hr, 2 hr, 4 hr, 6 hr, 12 hr and 24 hr. The 5 mn interval detections (the first top panel) that are indicated in grey on every other panels. The dots at the bottom of each panel indicate the average profiles processed by the algorithm. The larger the averaging interval is, the smaller the number of data (average profiles) is, the sparser the dots are.

6 Discussion and Conclusion

An objective method of PSC detection on backscatter profile is presented. The detection is based on the local increase in the profile variance produced by the presence of a PSC layer. The detection procedure consists in three steps. The first step consist of performing a stationarisation of the backscatter profiles. The second step involves the calculation of a maximum likelihoods. In the last step, the statistical efficiency of the PSC detection is estimated. The performances of the detection system are evaluated on simulated backscatter profiles that mimic typical characteristics of lidar profiles. The tests on simulated data show that PSC layers are reliably detected when they produce changes in variances greater than the background (i.e. PSC-free) variance. They also show that the dispersion of the estimated cloud bottom and top altitudes is found to be about 200 meters typically and that there is a systematic bias of about 300 m linked to the smoothing of the profiles.

After having been successfully tested on simulated data, the method is applied to real backscatter profiles measured above DDU station between March and October 2008. The results confirm the relevance of the detection algorithm. Series of PSC layers are detected during the austral winter and early spring (June, July, August and September). No PSC layer is detected during months when PSCs are not expected to form according to thermodynamical thresholds. The effect of temporal averaging has also been analysed. This averaging is often necessary when the lidar measurement time integration is very short. Its aim is to minimise the instrumental noise and hence maximise the signal-to-noise ratio. However the averaging degrades the temporal resolution and more importantly, if the temporal averaging far exceeds the inner variability time scale of the probed PSC layer, the measurements end up considering an overall optical smoothed equivalent of the cloud. The results suggest that the best compromise for PSC lidar detection at DDU is of the order of 1 hour.

There are other potential applications of this detection method presently applied to ground-based lidar profiles. A similar treatment could be applied to satellite lidar profiles. Since the optical signature of volcanic aerosol layers on lidar profiles is rather similar to the weak signal of optically small PSC, applying this method to the detection of volcanic layer appears straightforward (i.e. [David et al., 1998] and [David et al., 2009]). In the same way, the detection of other clouds (cirrus or noctulescent clouds [Von Cossart et al., 1996] or [Dubietis et al., 2010]) should also be possible with this approach. Finally, this could also be suited for the detection of biomass burning plumes or desert dust layers in tropospheric lidar profiles.

One limitation of the model is that it allows to detect only a single layer in a profile, precluding detection of superimposed PSC layers. Such improvement of the method requires new developments but no theoretical issues are to be overcome. As PSC backscattered signals depend on the lidar wavelength, the use of lidar profiles acquired with different wavelengths and a multivariate approach (one per wavelength) would allow to distinguish type of detected PSCs. A bayesian approach (see for example the development to variance shifts detection of [Hannart and Naveau, 2009]) could for instance be considered to tackle this new problem.

7 Appendices

7.1 Likelhood calculation

This annexe present the calculation which allows to infer the parameters of profiles. The first model, M_0, explained by (3.9) can be mathematically modelled by

$$M_0 : \forall z \in [z_1, z_n] \;\; P^*(z) \hookrightarrow \mathcal{N}(0, \sigma^{*2}_{out}). \tag{3.16}$$

This means that the distribution of the stationarized profile P^* is constant along the altitude range (i.e. $\forall z \in [z_1, z_n]$). Whereas the alternative model, M_1, explained by (3.10) is expressed by

$$M_1 : \begin{cases} \forall z \in [z_1, \tau_b[\cup[\tau_t, z_n] \;\; P^*(z) \hookrightarrow \mathcal{N}(0, \sigma^{*2}_{in}) \\ \forall z \in [\tau_b, \tau_t[\;\; P^*(z) \hookrightarrow \mathcal{N}(0, \sigma^{*2}_{in}) \end{cases}, \tag{3.17}$$

and means that two altitudes exist τ_b and τ_t which correspond to the bottom altitude and the top altitude of a hidden signal, within this altitudes the variance is supposed to be greater or equal to the variance outside.
Note that, if considering $\sigma^*_{in} = \sigma^*_{out}$ in equation (3.17), models from equation (3.16) turn out to be embedded in models from equation (3.17). To estimate the parameters of the model, the calculation of the likelihood maximum of distribution given by equation (3.17) is needed.
For all $z \in [z_1, z_n]$, the distribution function of $P^*(z)$ under M_1 is given by

$$\begin{aligned} f(P^*(z)|M_1) &= \frac{1}{\sigma^*_{out}\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^{*2}_{out}}[P^*(z)]^2) \;\; if \; z \in \; [z_1, \tau_b[\cup[\tau_t, z_n], \\ &= \frac{1}{\sigma^*_{in}\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^{*2}_{in}}[P^*(z)]^2) \;\; if \; z \in \; [\tau_b, \tau_t[, \end{aligned} \tag{3.18}$$

where $z_1 \leq ... \leq \tau_b \leq ... \leq \tau_t \leq ... \leq z_n$.
Assuming the random variables $P^*(z)_{z_1 \leq z_i \leq z_n}$ are independent, then, under M_1, the distribution of the vector $P^* = (P^*(z_1), ..., P^*(z_n))$ is given by

$$\begin{aligned} f(P^*|M_1) &= \prod_{z \notin [\tau_b, \tau_t[} \frac{1}{\sigma^*_{out}\sqrt{2\pi}} \exp(-\frac{[P^*(z)]^2}{2\sigma^{*2}_{out}}) \prod_{z \in [\tau_b, \tau_t[} \frac{1}{\sigma^*_{in}\sqrt{2\pi}} \exp(-\frac{[P^*(z)]^2}{2\sigma^{*2}_{in}}) \\ &= \left(\frac{1}{\sigma^*_{out}\sqrt{2\pi}}\right)^{n-\tau_t+\tau_b} \left(\frac{1}{\sigma^*_{in}\sqrt{2\pi}}\right)^{\tau_t-\tau_b} \prod_{z \notin [\tau_b, \tau_t[} \exp(-\frac{[P^*(z)]^2}{2\sigma^{*2}_{out}}) \prod_{z \in [\tau_b, \tau_t[} \exp(-\frac{[P^*(z)]^2}{2\sigma^{*2}_{in}}). \end{aligned} \tag{3.19}$$

The likelihood is then given by

$$\begin{aligned} &\mathcal{L}(\mathbf{z}; \sigma^*_{out}, \sigma^*_{in}, \tau_b, \tau_t) = \log(f(P^*|M_1)) \\ &= -n\log(\sqrt{2\pi}\sigma^*_{out}) + (\tau_t - \tau_b)\log\frac{\sigma^*_{out}}{\sigma^*_{in}} - \frac{1}{2}\left[\sum_{z \notin [\tau_b, \tau_t[} \frac{[P^*(z)]^2}{\sigma^{*2}_{out}} + \sum_{z \in [\tau_b, \tau_t[} \frac{[P^*(z)]^2}{\sigma^{*2}_{in}}\right]. \end{aligned} \tag{3.20}$$

For algorithmic performance, the previous likelihood can be written as

$$\begin{aligned} &\mathcal{L}(\mathbf{z}; \sigma^*_{out}, \sigma^*_{in}, \tau_b, \tau_t) = \\ &= -n\log(\sqrt{2\pi}\sigma^*_{out}) + (\tau_t - \tau_b)\log\frac{\sigma^*_{out}}{\sigma^*_{in}} - \frac{T}{2\sigma^*_{out}} + \frac{1}{2}(\frac{\sigma^*_{in} - \sigma^*_{out}}{\sigma^*_{in}\sigma^*_{out}}) \sum_{z \in [\tau_b, \tau_t[} [P^*(z)]^2. \end{aligned} \tag{3.21}$$

Where T is the total sum of squared $P^*(z)$ (i.e. $\sum_{z \in [z_1, z_n]} P^*(z)^2$). This last step allows to calculate only one of the two sums of equation (3.20).

The seek of the maximum of $\mathcal{L}(\mathbf{z}; \sigma^*_{out}, \sigma^*_{in}, \tau_b, \tau_t)$ regarding σ^*_{out}, σ^*_{in}, τ_b and τ_t is performed using a iterative method explained in Part 4.2.

Chapitre 4

Détection de pulses caractéristiques aux éruptions volcaniques : Application à des données de sulfate issues de carottes de glace.

Court résumé du chapitre :
Ce chapitre présente une méthode multivariée d'inférence de signaux de type éruptif x (voir Figure 4.1) appliquée à la détection de signaux volcaniques y_j de séries de sulfate extraites de carottes de glace. La série j de sulfate est modélisée par :

$$y_j(t) = f_j(t) + \beta_j x(t) + \epsilon_j(t), \tag{4.1}$$

où $x(t)$ est le signal éruptif, f_j représente la tendance du signal, β_j modélise les amplitudes respectives de chaque éruption et ϵ_j est l'erreur d'observation. Ce développement a requis la résolution des équations d'un filtre de Kalman dans un cadre non linéaire et non stationnaire, qui permet une décomposition simultanée des différentes composantes du modèle. Les difficultés de résolution d'un tel modèle viennent du caractère simultané et aléatoire du nombre, des temps d'occurrence et des amplitudes des signaux éruptifs cachés dans y_j, ainsi que de la prise en compte de tendances f_j propres à chaque série d'observation. Nous avons ensuite appliqué ce modèle à la détection de signaux volcaniques cachés dans les mesures de sulfate issues de carottes de glace forées au Groenland. Nous introduisons cette étude grâce à un préambule, suivi par un article soumis et sous révision dans le journal Computational Statistics & Data Analysis-Elsevier. Enfin nous présentons une étude des résidus afin de tester les hypothèses du modèle.

Plan du Chapitre 2

1. **Préambule**
2. **Introduction**
3. **Extraction Procedure**
4. **A simulation study**
5. **Application to Ice Core Data**
6. **Discussion and conclusion**
7. **Appendices**
8. **Validation a-posteriori de la méthode : caractéristiques des résidus**

1 Préambule

Les éruptions volcaniques ont un impact important sur le climat (e.g. [Castellano et al., 2004]). Le nuage de particules ou de précurseurs de particules est projeté dans la troposphère, parfois jusque dans la basse stratosphère pour les plus importantes éruptions . Les grandes éruptions volcaniques jouent un rôle dans l'évolution du climat, autant présent que passé. Il est donc nécessaire de pouvoir dater, estimer l'amplitude et évaluer l'incertitude de tels évènements passés pour comprendre l'évolution du climat jusqu'à nos jours et dans la même idée, pour forcer plus précisément les modèles numériques de simulation du climat passé.
Au fur et à mesure des années, la glace polaire piège à chaque saison les différents dépôts de matière transportée, notamment d'aérosols de sulfate d'origine stratosphérique pour les grandes éruptions, transportés jusqu'aux hautes latitudes par l'effet de la circulation générale stratosphérique. Le forage de carottes de glace et l'étude de leurs compositions chimiques permettent de mettre en évidence des marqueurs des différents grands évènements volcaniques remontant à plusieurs siècles. Nous disposons de plusieurs séries de données de sulfate issues de carottage de glace au Groenland, sur cinq sites géographiques différents couvrant la période allant de 1645 à 1980.

Les enregistrements des éruptions récentes montrent que la variation temporelle dans l'atmosphère des aérosols volcaniques peut être modélisée par un évènement soudain dont l'amplitude décroit ensuite pour finalement disparaitre. La dispersion et l'élimination totale des aérosols volcaniques dans l'atmosphère, notamment dans la stratosphère, peut durer plusieurs années. Á cela s'ajoute que chaque site peut présenter des évolutions particulières du fait de leur localisation et exposition. Les mesures montrent donc des évolutions lentes (comportements basses fréquences : une tendance due aux dépôts secs et humides de sulfate provenant de l'oxydation du diméthyle-sulphide DMS, voir l'article de [Castellano et al., 2004]) et des variations hautes fréquences (données bruitées par les variations minimes de distribution de sulfate et par la méthode de mesure) différents pour chaque site. Seuls sont susceptibles d'apparaitre de manière simultanée dans les séries, les évènements volcaniques qui ont été transportés massivement vers les hautes latitudes. L'extraction des différentes composantes du signal (tendance, bruit et éruptions volcaniques) à partir du signal initial fourni par les carottes de glace est l'objet de ce Chapitre.

Le modèle additif (4.1) permet de décrire les différentes caractéristiques des séries

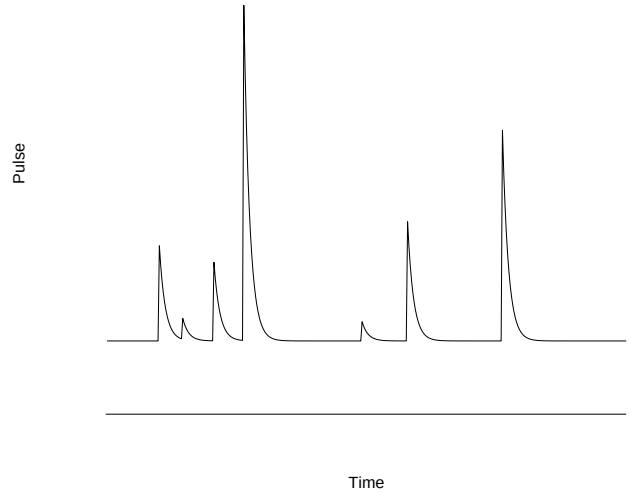

FIGURE 4.1 – Une simulation d'un signal éruptif x de l'équation (4.1) recherché dans les séries.

chronologiques de sulfate, à savoir, une tendance et un bruit propre, et la présence à intervalles aléatoires d'évènements volcaniques communs à chacune des séries mais ayant des amplitudes différentes d'une série à l'autre et d'un évènement à l'autre. Ce problème est résolu grâce au développement d'un filtre de Kalman non linéaire multivarié.
Certains articles de la littérature (par exemple [Castellano et al., 2004]) décrivent le fait que les amplitudes des volcans (mesurée par la quantité de sulfate produite) suivent une loi lognormale. Or notre modèle suppose que cette distribution suit une loi normale. Ce choix a été testé en faisant tourner notre modèle à la fois sur le signal initial y, et sur sa $\log$-transformation $\log(y)$. En effet, la tendance des séries de sulfate que nous étudions étant très faible (i.e. le rapport signal sur bruit proche de un), on a considéré, en première approche grossière que si y suit une loi log-Normale alors $\log(y)$ suit une loi Normale. Les détections des instants d'éruptions étant identiques quel que soit le signal d'entrée utilisé, nous avons choisi de présenter les résultats sur le signal initial y, car cela permet d'exploiter la composante *tendance*, f, de notre modèle (4.1), ce qui n'aurait pas été le cas si nous avions travaillé sur la transformation $\log(y)$. Cette composante représente le comportement basse fréquence de la distribution temporelle de sulfate dans l'atmosphère, et permet ainsi d'évaluer l'évolution de la concentration de sulfate dans l'atmosphère.

Les résultats sur l'application ont permis de détecter des éruptions volcaniques bien connues, tels que le Laki, en 1783, ou l'éruption du Katmai-Novarupta en 1912. L'intérêt de cette méthode réside essentiellement dans la détection d'éruptions secondaires, plus difficilement détectables. Le tableau 4.1 énumère les éruptions détectées par notre méthode, leurs amplitudes et l'incertitude que l'on a sur ces détections.
Les perspectives qu'offre ce modèle, au delà de notre application sont assez multiples, par exemple, un tel modèle permettrait la détection de pics de production de dioxyde d'azote (NO_2) suite à des éclairs qui prennent des formes similaires aux éruptions volcaniques sur des périodes de temps bien plus courtes (voir les articles de [Boersma et al., 2005], [Fraser et al., 2007]). Certaines améliorations du modèle permettraient également de prendre en compte des bruits qui ne soient pas indépendants ou encore des instants de détection qui ne soient pas simultanés, mais pouvant légèrement varier d'une série à l'autre, ce qui rendrait le modèle plus souple. Ce qui permettrait entre autre de palier à certains défauts de mesure liés à la manière dont les couches de glace se forment dans les carottes. Le problème de log-normalité posé précédemment nous amène aussi à considérer le développement futur du modèle suivant :

$$y_j(t) = f_j(t) + \beta_j \exp(x(t)) + \epsilon_j(t), \tag{4.2}$$

qui modeliserait des amplitudes d'éruptions distribuées de manière log-normale.
L'article présenté dans ce chapitre est organisé comme suit. Une première partie d'introduction décrit l'enjeu des éruptions volcaniques pour l'étude du climat et explique brièvement les choix faits dans cette étude au regard du comportement des signaux volcaniques. Une seconde partie définit le modèle de détection et le choix du modèle d'état. Le modèle est testé sur des données simulées dans la partie suivante, et ensuite appliqué aux données de sulfate issues des carottes de glace. Enfin, nous concluons en expliquant les résultats et en discutant le modèle de détection présenté. Dans l'annexe est détaillé l'ensemble des calculs du filtre de Kalman et son adaptation au non linéarité des signaux volcaniques, ayant permis la réalisation de ce système de décomposition de signal.

Article #2 : doi :10.1016/j.csda.2012.01.024 - ***Computational Statistics and Data Analysis.***

Extracting Common Pulse-Like Signals from Multiple Ice Core Time Series

Julien Gazeaux[1], Deborah Batista[2], Caspar M. Ammann[3], Philippe Naveau[4], Cyrille Jégat[5], Chaochao Gao[6] [1]Laboratoire Atmosphère Milieux et Observations Spatiales, IPSL-CNRS, Paris, France

[2]Department of Mathematical Sciences, University of Colorado Denver, Denver, CO, USA

[3]Climate and Global Dynamics Division, National Center for Atmospheric Research, Boulder, CO, USA

[4]Laboratoire des Sciences du Climat et l'Environnement, IPSL-CNRS, Gif-sur-Yvette, France

[5]Ecole Nationale Supérieure des Mines de Paris, Paris, France

[6]Department of Environmental Sciences, Rutgers University, New Brunswick, NJ, USA

abstract

To understand the nature and cause of natural climate variability, it is important to possess an accurate estimate of past climate forcings. Direct measurements that are reliable only exist for the past few decades. Therefore knowledge of prior variations has to be established based on indirect information derived from natural archives. The challenge has always been to find a strictly objective method that can identify volcanic events and offer sound amplitude estimates in these noisy records. An automatic procedure is introduced here to estimate the magnitude of strong, but short-lived, volcanic signals from a suite of polar ice core series. Rather than treating records from individual ice cores separately and then averaging their respective magnitudes, our extraction algorithm jointly handles multiple time series to identify their common, but hidden, volcanic pulses. The statistical procedure is based on a multivariate multi-state space model. Exploiting the joint fluctuations, it provides an accurate estimator of the timing, peak magnitude and

duration of individual pulse-like deposition events within a set of different series. This ensures a more effective separation of the real signals from spurious noise that can occur in any individual time series, and thus a higher sensitivity to identify smaller scale events. At the same time, it provides a measure of confidence through the posterior probability for each pulse-like event, indicating how well a pulse can be recognized against the background noise. The flexibility and robustness of our approach, as well as important underlying assumptions and remaining limitations, are discussed by applying our method to first simulated and then real world ice core *Signal Extraction, Multiprocess Kalman Filter, volcanic eruptions, pulse-like signals, climate forcing*

2 Introduction

2.1 Records of volcanic eruptions

Explosive volcanic eruptions are extreme events that can inject large amounts of sulfur-baring gases into the stratosphere. There, the gases are converted into small sulfuric acid droplets that spread and blanket the planet with a light "dry haze" [Lamb, 1970] that scatters and reflects sunlight. The resulting reduction of solar radiation to the surface can have a strong, albeit short-lived, impact on the climate ranging from several months to a few years [Robock, 2000]. Volcanic cooling was found in numerous climate time series, instrumental and proxy alike [Bradley, 1988, Briffa et al., 1998, Crowley, 2000, Jones et al., 2003, Robock and J.Mao, 1995]. Through a large event, or clustering of smaller eruptions, volcanic forcing is thought to be one of the primary factors affecting decadal to century scale evolution of climate [Ammann et al., 2007, Crowley, 2000, Hegerl et al., 2003, 2006]. It is therefore important to have a good quantitative estimate of these perturbations through time.

Volcanic forcing histories can be estimated from a host of sources, such as volcanological records [Newhall and Self, 1982, Simkin and Siebert, 1994], instrumental records [Sato et al., 1993, Stothers, 1996b, Ammann et al., 2003], observational information of visible perturbations of atmosphere or the ground [Lamb, 1970], or astronomical observations [Keen, 2001]. However, the probably most reliable records that are most consistent in time come from polar ice core series [Gao et al., 2008, Hammer, 1977, Robock and Free, 1995, Zielinski et al., 1994], where volcanic acid or sulfate spikes can be identified within individual snow and ice layers of the generally pristine environments of the ice caps [e.g., see Robock, 2000, Zielinski, 2000, for discussion].

Various techniques have been used to recognize these volcanic deposits, either using

electrical conductivity changes to identify the variations in acidity [Hammer, 1977], or more recently through direct measurements of sulfate at very high resolutions throughout the ice cores [Zielinski et al., 1994, Clausen et al., 1997, Palmer et al., 2001, Castellano et al., 2005, Kurbatov et al., 2006]. The advantage of using the polar ice sheets as an archive for individual volcanic events is that they preserve the climatically all-important sulfate. If sulfate can be found at multiple locations, then it is highly likely it was transported through the stratosphere, and thus was climatically "active" [Zielinski et al., 1994, Clausen et al., 1997]. (In contrast, tropospherically transported sulfate is too short lived in the atmosphere and thus is unlikely to have significant climatic effects.) Deposition can happen both through slow and evenly distributed dry deposition, or through more event-like wet deposition associated with storm systems. The volcanic sulfate signals that can be found at various ice-core locations, therefore, represent a spatial sample of the large-scale deposition. Areas with generally more storm events also commonly exhibit higher sulfate deposition rates. Thus, while each event will likely have some unique weather-related depostion features, there is an underlying spatial pattern that reflects the climatological deposition rates [Mosley-Thompson et al., 1993, Gao et al., 2007]. Based on a suite of cores, and thus multiple samples for each event, one can determine what the timing and the flux of sulfate was to the ice sheet, which in turn can be used to estimate the amount of sulfate and thus forcing.

As with all indirect information, using polar ice cores also involves some inherent difficulties. Exact dating of individual ice layers, a pre-requisite for core-to-core comparisons, is more problematic than in biological records, such as tree rings, where time progresses without interuptions. This continuity is not always guaranteed for the small diameter (< 15cm) ice cores because the possibility for stratigraphic disturbance exists. Wind can under some circumstances errace snow layers in such small areas ; sometimes it can accumulate more snow, which then forms a false "annual" band. Therefore, the ice coring community has been using characteristic time markers, and in particular a few of the largest volcanic events, as cross-dating hinge-points. Although generally defendable from a physical perspective, this approach could potentially introduce some biases. Exploiting relative time intervals of volcanic signals between these marker events as well as inclusion of other prior knowledge have been corner stones of "expert-based" ice core compillations for volcanic forcing reconstructions [Crowley, 2000, Ammann and Naveau, 2003, Ammann et al., 2007, Gao et al., 2006]. A more objective method is desirable. Here we develop the foundation for such a method, but we have to rely on the assumption that the available chronologies are perfectly dated (synchronized), a condition that would have to be assessed in more detail for the full set of existing high-resolution ice core records.

2.2 The statistical problem

Statistically, volcanic perturbations can be viewed as pulse-like events, i.e. short and intense deviations from the climatological (interannual) noise and some underlying longerterm variation. The common procedures to identify and then quantify the volcanic signal has been by applying an evolving mean to the individual time series and then selecting signals that pass a certain threshold of noise around this mean [Gao et al., 2008, Robock and Free, 1995]. However, classical statistical tools based on averages, variances or projections are not well adapted to capture the true characteristics of the rapid and sharp features of the volcanic eruption and deposition process [Naveau and Ammann, 2005].

Recently, more modern statistical tools using state space models have been used to improve on the identification and quantification process of volcanic events (and thus enhance objectivity) in individual time series [Naveau and Ammann, 2005]. Identified events could then be averaged across the different series. While large events were generally easy to recognize, the extraction proved much more difficult for small events that are obscured by the background noise [Naveau et al., 2003]. If chronologies of ice cores were synchronized, then modern statistical tools could further exploit the common deposition structure across multiple ice cores, and thus the threshold to recognize events could be lowered significantly. This could make volcanic forcing series more reliable. We build on earlier work [Naveau et al., 2003, Naveau and Ammann, 2005] to develop a statistically sound volcanic extraction process that uses the joint information across a series of ice cores. The geophysical motivation is centered on volcanic sulfate deposits in ice core time series, but techniques developed in this article can equally be used in other applications where large amplitude pulses that are superimposed on slowly changing trends need to be recognized across multiple noisy time series.

One of the difficulties of statistical modeling in the multivariate framework resides in the estimate of a nonlinear and non-Gaussian hidden signal common to all of the time series. For example, in the case of volcanoes, a big eruption will have a strong signal, but the relative magnitude might differ substantially between each of the ice core data sets, and the other components of the series could be quite heterogeneous. From a statistical point of view, a global solution to estimate the parameters is preferable because it reduces the propagation of errors. To solve these problems of extraction, we propose a *multivariate multi-state space model* which integrates the various components (forcing, trends, and noises) in a global mathematical formulation.

We discuss the general concept and properties of our extraction model in Section 3. Then, in Section 4, simulations with a known pulse process embedded in series with

different noise characteristics are used to assess the performance of our extraction method. The method is then applied to real ice core data proxies in Section 5. In Section 6, we conclude this article with a brief summary and discussion of the advantages, limitations and possible extensions of our extraction algorithm.

3 Extraction Procedure

3.1 State Space Modeling

State space models have become a practical and powerful tool to model dynamic and complex systems. Closely related to the Kalman filter, they has been used in a wide range of disciplines : biology, economics, engineering, and statistics [see Guo et al., 1999]. The fundamental idea of the state space model is that the observed data is linearly dependent on latent variables of interest that vary in time. Mathematically, the observed data are governed by two equations, known as the *observational* and *system equations*. In our case, the observational equation expresses itself as a linear combination of three variables (common forcing, trends, and noise), while the system equations represent the temporal dynamics of the underlying hidden processes. The statistical problem is to deduce the behavior of hidden variables of the pulse-like events from the observed data. Before discussing the aspects of our multivariate approach, we must introduce some notation and clarify our working hypotheses.

3.2 Our model

Suppose we observe J time series over the same time length, say T, and with the same temporal resolution. Each time series is denoted $y_j(t)$. We also assume that each of these time series is affected by a similar pulse-like forcing, say $x(t)$, that is unobserved and has to be estimated. This forcing corresponds to abrupt events and therefore is nonlinear and non-Gaussian. Our first equation explains how the three elements of our statistical model (trends, cycles, pulse-like events and noises) interact

$$y_j(t) = \beta_j\, x(t) + f_j(t) + \epsilon_j(t) \quad \begin{array}{l}\text{for } j = 1, \ldots, J \\ \text{and } t = 1, \ldots, T.\end{array} \tag{4.3}$$

The hidden, but common, pulse-like signal is represented by the random variable $x(t)$. The scalar β_j can be viewed as a scaling factor that reflects the impact of $x(t)$ on the j-th time series. In case of the occurrence of a pulse (i.e. $I_t = 1$), the random variable $v(t)$

is driven by a Gaussian distribution (the equation (4.5) is introduced below). That means that each event has its particular amplitude. The second component $f_j(t)$ corresponds to a smooth trend. The last term, $\epsilon_j(t)$, is simply a background iid Gaussian random noise process centered about zero with standard deviation σ_j. The different noises in Equation (4.3) are assumed to be independent.

The two main differences of our hidden signal $x(t)$ with classical regression models come from its pulse-like nature and its short term memory. To obey this constraint, we construct $x(t)$ as an autoregressive model defined by

$$x(t) = \alpha\, x(t-1) + v(t), \text{ for } t = 1, \ldots, T, \tag{4.4}$$

where $|\alpha| < 1$ is an unknown constant representing the decaying volcanic aerosol removal from the stratosphere and $v(t)$ corresponds to an iid random sequence and we set $x(0) = 0$. To create a pulse like effect, we impose that the iid random sequence $v(t)$ follows a mixture of a normal random variables

$$v(t) = \begin{cases} N(\mu_v, \sigma_v^2) & \text{, if } \quad I_t = 1, \\ 0 & \text{, if } \quad I_t = 0. \end{cases}, \tag{4.5}$$

where $N(\mu_v, \sigma_v^2)$ represents a Gaussian variable with mean μ_v and standard deviation σ_v. In Equation (4.5), I_t is a sequence of iid Bernoulli random variables, whose parameter $\pi = \Pr[I_t = 1]$ denotes the probability of observing a pulse-like event. The random variable $v(t)$ corresponds to the strength associated with a rare event. In contrast, $v(t)$ is set to zero if I_t equals to zero. Equation (4.4) allows for a short lived temporal effect of such a forcing. Despite its low number of parameters (π, α, μ_v, β_j, σ_j), the additive model defined by Equation (4.3) with this hidden dynamical structure (4.4) and its pulse-like nature defined by (4.5) offers enough flexibility to mimic pulse-like events behaviors at an annual scale.

The trends f_j in Equation (4.3) are modeled by cubic smoothing splines represented in state space form (see [Wahba, 1978] and [Wecker and Ansley, 1983]). This representation allows to express a *smooth* function (i.e. a function of which a fixed number m of derivatives are continous) as a sum of its weigthed derivatives and a Wiener process. By choosing $m = 2$ trends $f_j(t)$ can be expressed as functions of its first derivatives,

$$\mathbf{F_j}(t) = B\mathbf{F_j}(t-1) + \mathbf{E}_{f_j}(t),$$

where $\mathbf{F_j}(t) = (f_j(t), f_j^{(1)}(t))$, $B[i,k] = 1/(k-i)!$ for $k \geq i$ or zero otherwise. The

two-dimensional vector $\mathbf{E}_{f_j}(t)$ represents a zero mean Gaussian vector with covariance elements $\lambda_j\sigma_j^2/[(i+k-1)(i-1)!(k-1)!]$ where λ_j denotes the smoothing parameter. This spline representation allows to model non parametric trend, based on series developments similar to Taylor's. The trend is calculated using a polynomial regression and a random residues modelled by a Wiener process (see [Abramowitz and Stegun, 1970] and [Stark and Woods, 2002]).

With these notations, it is possible to combine $x(t)$, $f_j(t)$, and their associated noises, and thus to rewrite equations (4.3-4.5) in matrix form. With the state vector $X_t = (v(t-1), x(t), \mathbf{F_1}(t), \ldots, \mathbf{F_J}(t))^T$ We can define

$$Y_t = HX_t + E_t, \tag{4.6}$$

where the temporal dynamics is then described by another matrix equation

$$X_t = \Phi X_{t-1} + E_t^*. \tag{4.7}$$

The matrices H and Φ and the random vectors E_t and E_t^* have explicit (but complex) forms that are given in the Appendix.

A rich literature [Guo et al., 1999, Shepard, 1994] exists to estimate parameters of the state space models. Such techniques are closely related to statistical data assimilation schemes. In Gaussian state space models, the Kalman filter provides an optimal recursive estimate of $x(t)$ from observations $Y_t = (y_1(t), \ldots, y_J(t))$. Unfortunately, the nature of the pulse-like events (the mixture of distribution) implies that the overall assumption of normality is not satisfied (see Equation 4.5). To solve this problem, we drew inspiration from original work of [Guo et al., 1999] who offered a variation of the Kalman filter. The principal idea is to approximate the distribution of the mixture of normals by a normal distribution whose first two moments are identical to that of the mixture. The details of this technique within the univariate framework can also be found in [Naveau et al., 2003]. When the last evaluations of this modified Kalman filter are found, then a sequential backward algorithm is applied [Guo et al., 1999].

4 A simulation study

We show in Figure 4.2 three simulated times series, each with a different trend, a background variation of local noise and the common pulse-like signal with its site specific scale (the simulation series were made with $\mu_v = 3.5$, $\sigma_v = 2.63$, $\alpha = 0.7$, and $\pi = 0.3$,

as parameters of Equations (4.4) and (4.5)). The included pulse-process is shown in the two bottom panels of Figure 4.2. In the top panel of Figure 4.2, the time series $y_1(t)$ combines a discernible cycle (a sinusoidal trend with a constant level shift defined by $f_1(t) = 10 + 15\sin(2\pi((t-1)/90))$ with noise and the pulse-like forcing. In this data set, most pulse-like events are visually identifiable because the noise level is low compared to the pulse amplitude. The middle panel shows a more noisy time series $y_2(t)$ with a linear trend ($f_2(t) = 0.5t$). Here, finding pulse-like events represents already a more difficult challenge. For instance, the small pulses at the beginning of this time series (see bottom panels of Figure 4.2), clearly visible in $y_1(t)$, are not easily distinguishable in time series $y_2(t)$. Although the series in the bottom panel of Figure 4.2 contains no trend ($f_3(t) = 0$), this series $y_3(t)$ is characterized by the lowest multiplicative factor ($\beta_3 = 7.5$) for the pulse-like events among the three simulated time series, i.e. the common underlying signal in $y_3(t)$ is less apparent in the large noise.

Figure 4.3 shows the results in the multivariate case (top panel) as well as for each of the three individual series if they would be treated separately. Each panel compares the identified events (solid line) with the timing of the true events that were embedded as hidden pulses in all three series (gray bars). Overall, the multivariate model was effective in identifying the hidden pulses and the joint extraction is capable of highlighting features that were not detected by the individual analyses, e.g. the double peak just before $t = 400$.

The increasing noise in series $y_2(t)$, and particularly $y_3(t)$, obviously impacts the extraction process. The shortcomings express themselves both in the less accurate identification (less) of the imposed events as well as in the rather volatile event magnitude.

This is illustrated through a scatter plot (Figure 4.4) for each identified event, where the magnitude of the true (hidden) process (x) is compared against the estimates ($\hat{x}$). The graph highlights that the multivariate algorithm is able to estimate most accurately (up to a multiplicative constant) the amplitudes of the hidden pulses. Also the extraction from the low-noise series $y_1(t)$ was successful and contains most of the events. The overall estimate of the amplitude across all cases is more accurate in the multi-variate case, despite two rather unfavorable additional series $y_2(t)$ and $y_3(t)$.

The extraction procedure not only quantifies the pulse-like events but also offers the full information about the underlying trends. Figure 4.5 shows that the three trends f_j of Equation (4.3) are well captured.

5 Application to Ice Core Data

In the real world, the most reliable records of volcanic pulses come from ice cores.

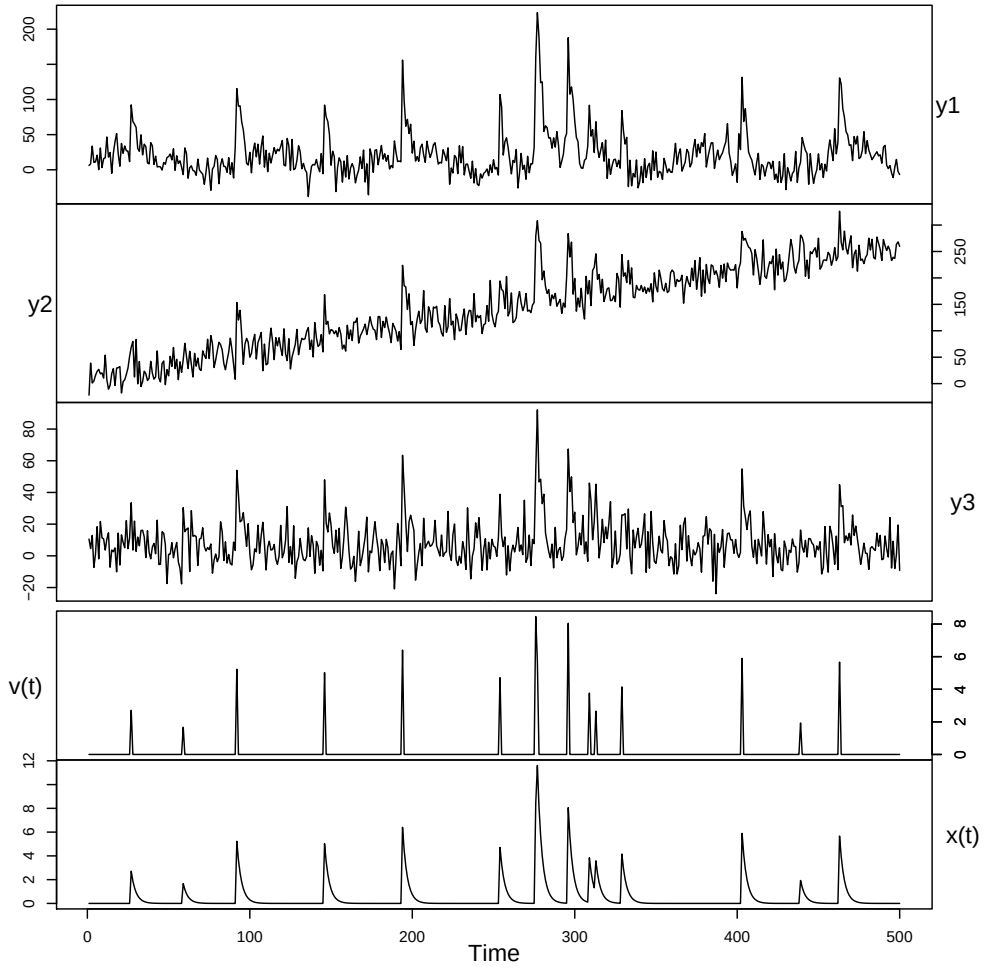

FIGURE 4.2 – Simulated data from Equation (4.3) with $J = 3$. All series represent samples over a time span of 500 years and were simulated with the following parameter setting : the standard deviations observation noises : $(\sigma_1, \sigma_2, \sigma_3) = (15, 20, 10)$, the parameters of pulse amplitudes : $(\beta_1, \beta_2, \beta_3) = (20, 15, 7.5)$, the pulse occurence probability : $\pi = 0.03$, the Auto-Regression parameter of x : $\alpha = 0.7$, the common mean pulse amplitude of v : $\mu_v = 3.5$, the standard deviation of pulse event amplitude : $\sigma_v = 2.63$. The two bottom panels represent the simulated pulse-like time series hidden in the three time series obtained from equations (4.4) and (4.5) with $\mu_v = 3.5$, $\sigma_v = 2.63$ and $\pi = 0.03$.

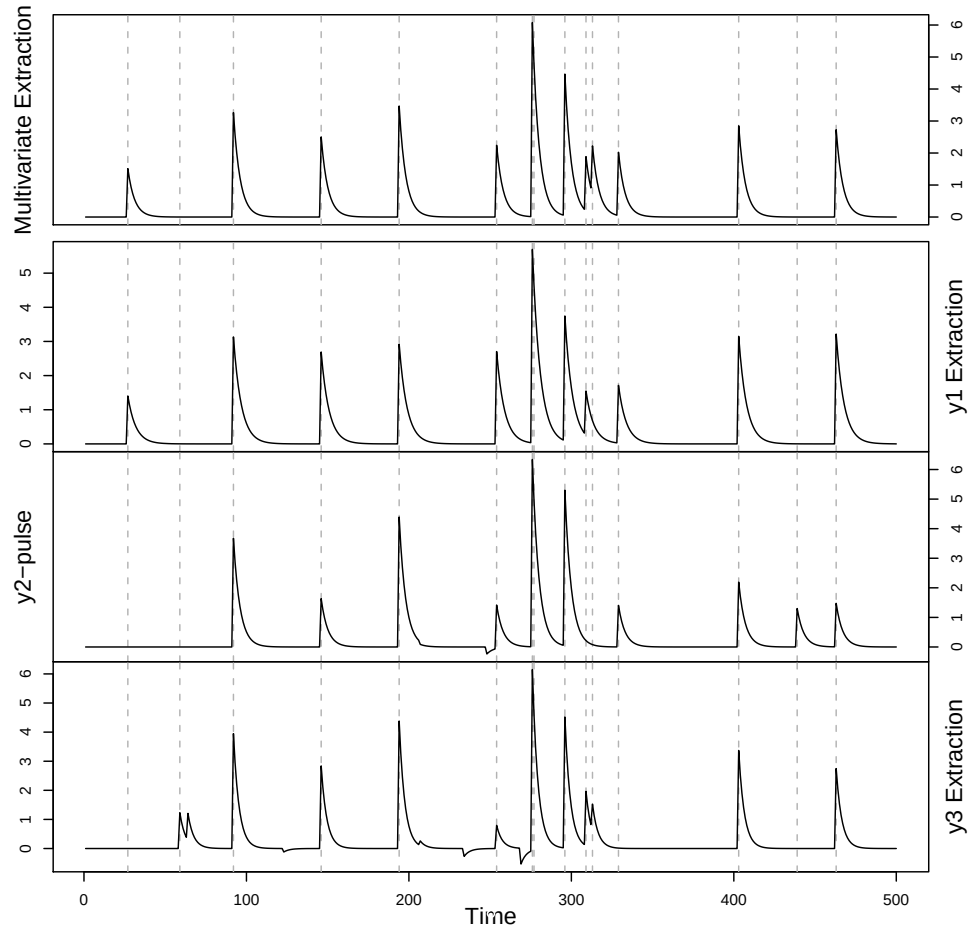

FIGURE 4.3 – Estimated pulse-like amplitudes. The top panel corresponds to the *multivariate* extraction and the other three panels represent the *univariate* extraction applied to individually to each time series from Figure 4.2, respectively y_1, y_2 and y_3. Note that the multivariate extraction dismissed the "negative" spurious events detected on the 2nd and 3rd series. Note also that the multivariate extraction allows to detect more actual pulse like events than the different univariate cases.

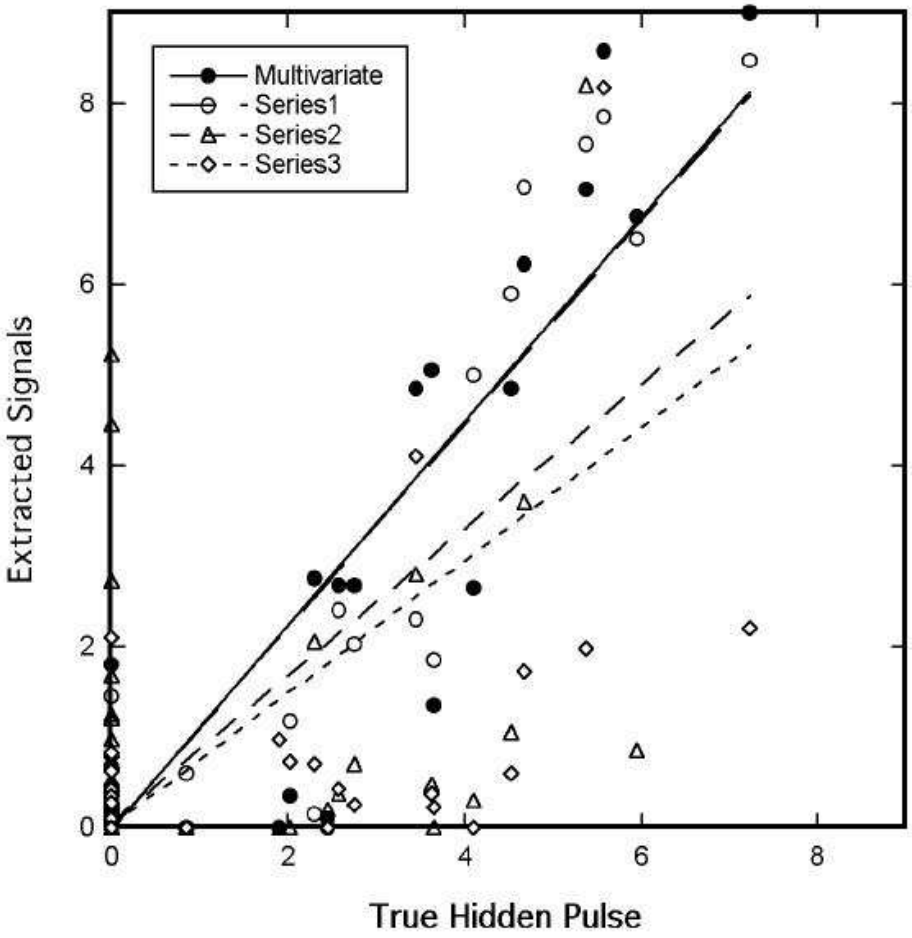

FIGURE 4.4 – The x-axis represents the standardized hidden $x(t)$ and the y-axis corresponds to our estimated standardized $\hat{x}_t$ from the data displayed in Figure 4.2.Black circles corresponding to the multivariate extraction better estimate amplitudes of the pulse like events than the different univariate cases.

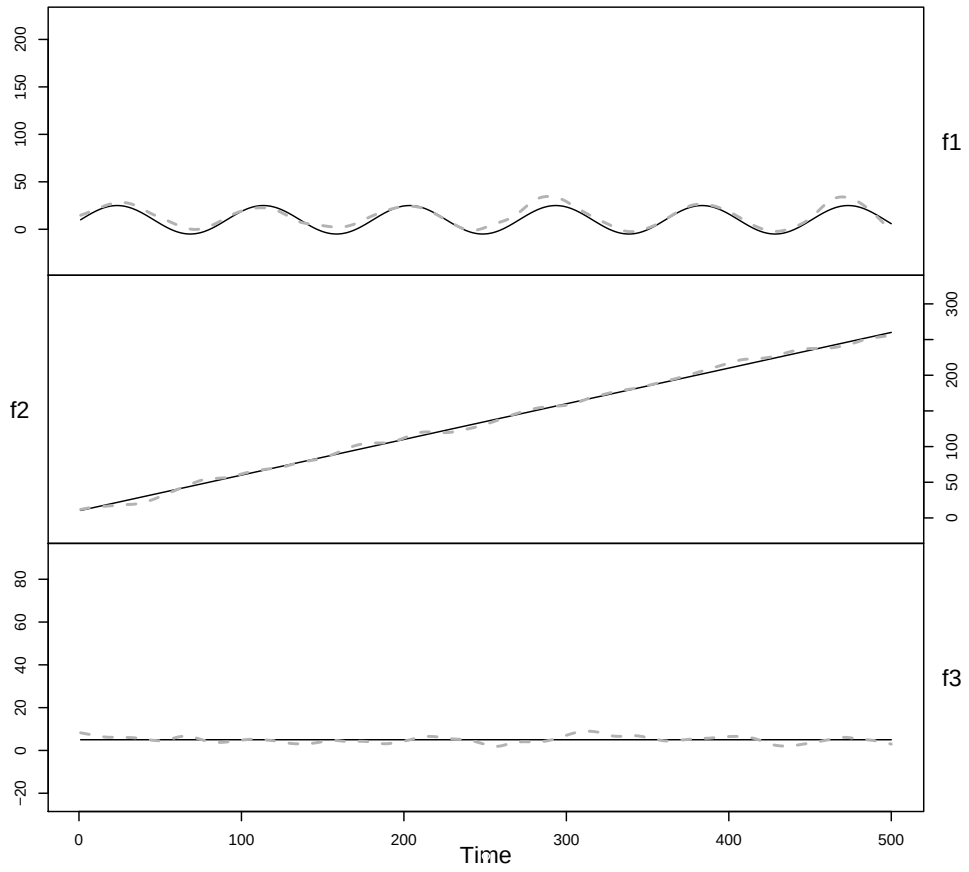

FIGURE 4.5 – The solid black curves represent the hidden trend f_j and the dotted lines correspond to our estimated trend $\hat{f}_j$ for each of the time series displayed in Figure 4.2.

The idealized experiment shown above does not represent the true levels of noise and core-to-core differences in the the deposition. Therefore, to test our method with realistic data, we now apply the multi-variate extraction to five selected series from Greenland covering the period from 1645 to 1980 at annual resolution. As indicated above by the simulation results, doing a multivariate extraction rather than separate individual analysis that subsequently gets averaged has benefits for both recognizing particularly small events against the varying noise of different cores, and estimate more reliably the amplitude of volcanic pulses across Greenland where individual cores have substantially different absolute signals. Because we apply a joint-signal extraction through the multi-variate state-space model, we obtain a unitless, joint volcanic pulse history that is based on the mean contributions from individual ice cores. This unitless series can then be calibrated against known (i.e. measured) volcanic deposition or forcing, a substantial improvement over previous methods where individual records had to be calibrated based on a few noisy events, and to obtain an overall series, these noisy estimates had to be averaged.

Figure 4.6 shows five ice core records from Greenland with the identified trends from the multi-variate extraction procedure. Differences in series, their trends as well as the variance of the sulfate deposition are roughly representative for the full set of polar ice cores.

Table 4.1 and 4.7 show the extracted pulses (top panel) and their associated posterior probabilities (middle panel) resulting from the multi-variate state-space model. Five high-probability, about three intermediate-probability and several low-probability events are recognized across the series. While large events are often identified with high confidence, applying the extraction procedure to individual ice cores would include more, but often erroneous spikes in the list. The bottom panel of Figure 4.7 illustrates this fact with the extracted signal on the first of the ice core series. A whole series of "high-probability" events are "found" throughout the 20th century. These spikes, however, don't have any counterparts in the other series, and thus, despite the fact that they are large and recognized with high confidence in the individual record of series one, the *joint* likelihood is much lower. At the same time, more of the small volcanic inputs are recognized in the joint extraction with higher confidence, a capability that cannot be achieved at the individual level where small events tend to get overwhelmed by the background noise. The joint extraction reduces the noise and therefore is more sensitive towards small events.

Looking at the magnitude of events, one particular deposition event in the year 1783 is immediately recognizable across the five series. It is the result of the well documented, large and intensive Laki eruption sequence in Iceland ([Thordarson and Self, 2003]).This eruption quite likely did lead to a substantial non-stratospheric transport of sulfate to-

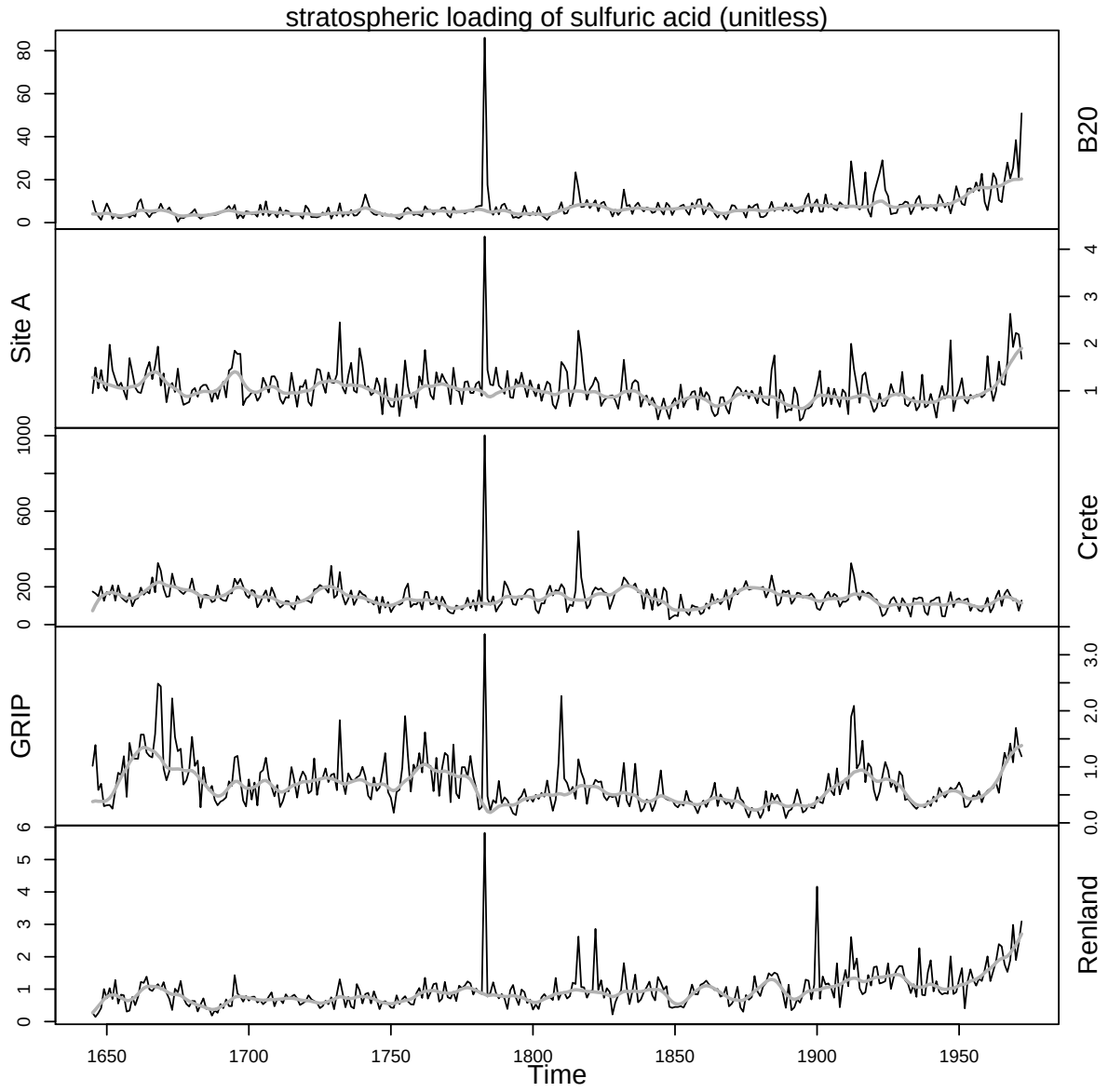

FIGURE 4.6 – Five ice core records of sulfate deposits from Greenland covering the period from 1645 to 1980 at annual resolution. with the estimated trends obtained from our multivariate extraction.

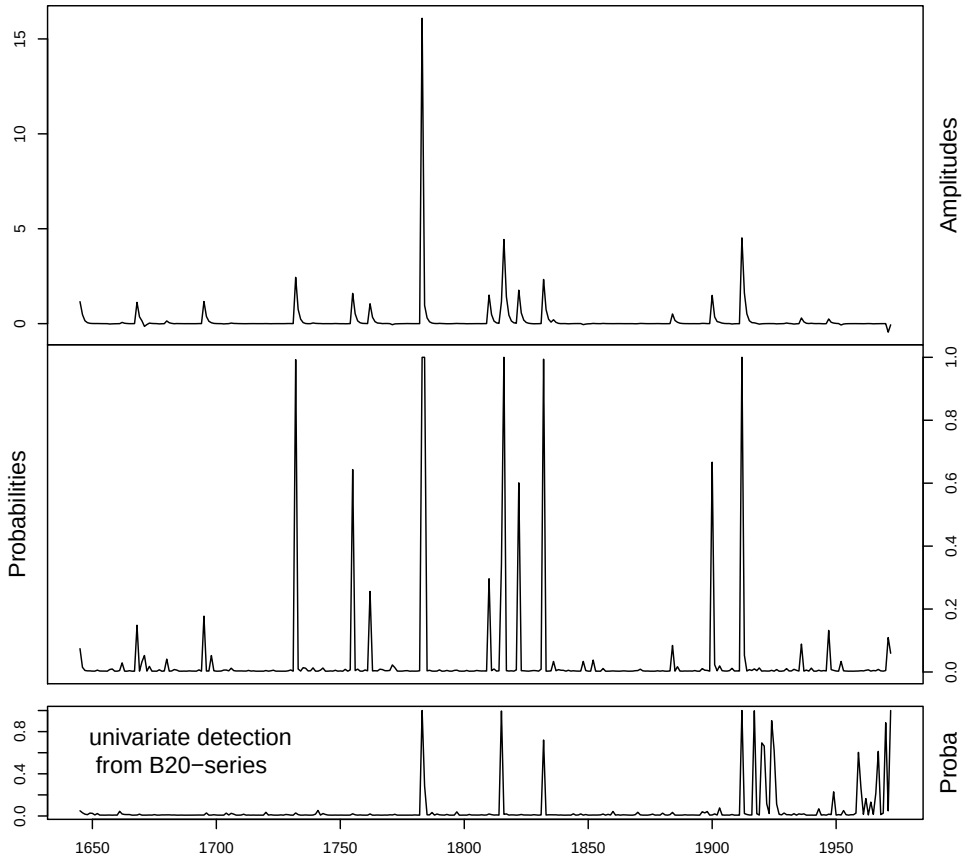

FIGURE 4.7 – Estimated magnitudes and associated event probabilities extracted from the five ice cores using the multivariate extraction approach. The bottom panel illustrates the erroneous spikes extracted using a univariate procedure throughout the 20th century.

wards nearby Greenland, causing the exceptional volcanic signal. Therefore, compared to low-latitude events whose sulfate aerosol get spread in the stratopshere across both hemispheres, this eruption appears disproportionally large, despite the fact that its regional effects were tremendous [Stothers, 1996a].

In contrast, the famous tropical eruption of Tambora of 1815 was recognized by the extraction with a signal less than one third of the 1783 Laki deposit, but nevertheless it was identified with high confidence (high posterior probability). Given the global spread of its aerosol, and no chance for any tropospheric transport of the sulfate, the stratospheric mass loading of Tambora must have been very large. Yet, it has to be calibrated differently than the Laki signal to obtain a global stratospheric mass. This difference between quasi-local high-latitude events and tropical eruptions represents one of the difficulties in the interpretation of polar-based ice core records of volcanic forcing. While corrections can easily be applied for these well-known events, lesser known, or even unknown, eruptions require more detailed analyses, such as the evaluation of other, the timing and deposition of other chemical species that could indicate local transport([Clausen et al., 1997]), or synchronous deposition in Antarctica and Greenland which would point to a tropical source([Palais et al., 1992], [Langway et al., 1995], [Ammann and Naveau, 2003]). But particularly for small events, this distinction is very difficult ([Crowley, 2000]).

The fifteen events listed in Table 1 contain some well-known events as well as some clearly unknown signals. We list the likely source volcanoes, but should not forget to consult the posterior-probability to evaluate if these events are to be regarded as robust.

To convert the volcanic pulse history into an estimate of a forcing, the amplitudes of the unitless ice core-extracted volcanic pulses need to be scaled to units such as "stratospheric loading of sulfuric acid" [Gao et al., 2008, Zielinski et al., 1994] or radiative forcing ([Hansen and Coauthors, 2002], and [Wigley et al., 2005]). The individual ice cores series don't contain local volcanic pulse signals that are directly comparable because of spatial differences in volcanic sulfate deposition [Gao et al., 2007]. Therefore, the output from the extraction procedure actually provides only the relative magnitudes of events. This series needs to be *calibrated* against other information, such as the conversion of the mean sulfate fluxes at a particular site to the stratospheric loading over a known period or by simply using a reference event (e.g. Pinatubo or El Chichon) [see Gao et al., 2008, for discussion].

As discussed for the simulated cases above, real world records exhibit often a combination of variations coming from an evolving trend and background noise in addition to the volcanic pulses. Some of these variations can be simple noise, others could potentially be interesting climatically. However, particularly with regard to sulfate deposition,

real world records sometimes suffer from a distinct, fundamental change over the last century as human fossil fuel burning has artificially released large amounts of sulfur into the atmosphere. Thus, most Northern Hemisphere records exhibit a systematic increase in the background deposition [Mayewski et al., 1986, Neftel et al., 1985]. Sulfate records from the 20th century should therefore be analyzed and interpreted with caution. Looking at individual records (Figures 4.6 and 4.7), a clear increase in the background noise as well as in the variance can be seen. The danger of erroneously identifying some of these anomalies as volcanic spikes exists in the individual extraction (Figure 4.7b), but the joint extraction is clearly less sensitive to this. Nevertheless, all of 20th century signals should be interpreted with caution.

In summary, compared to individual extraction of volcanic signal for each series, the joint extraction offers a more robust identification of events against noise, and the sensitivity for capturing small eruptions is increased. This benefit is already recognizable in this small sample of five ice core records from Greenland. In a future study, we will apply this technique to the full set of volcanic series from both Greenland and Antarctica to establish a new volcanic sulfate history that can be added to the currently existing series of [Crowley and Kim, 1999], [Ammann et al., 2007], [Gao et al., 2008] and [Wastegard and Davies, 2009].

Before discussing this work, we want to add a few comments about the fitting of the model with the actual series. The first question underlying this application is about the use of a Gaussian distribution in Equation (4.5). The choice of Gaussian distribution could implies the occurrence of "negative" volcanic events. This could happen when the mean μ_v of $v(t)$ is small with respect to the variance σ_v. In the simulation shown in Figure 4.2 we choose μ_v large enough to avoid that to happen. In the application, the parameter μ_v appears to be large enough with respect to σ_v.
Another related comment is needed. Volcanic eruptions are sometimes assumed to be driven by lognormal distribution instead of Gaussian (e.g. [Castellano et al., 2004]). To test this assumption, we run our model on the log transformation of the signal $y_j(t)$: $\log(y_j(t))$. This transformation allows roughly to change the supposed log-normal distribution of the amplitudes of the events into a normal distribution. Results of this study show exactly the same detected events as when running the model directly on $y_j(t)$. For our data there is no loss on using directly $y_j(t)$ as input of our model. This illustrates the robustness of our method (i.e. the way the algorithm acts when the Gaussian hypothesis is not guaranteed). This can be explained by the fact that our method is based on the existence of the first and second orders of the distribution (i.e. the mean and the variance).

Year	Ampl	Proba	Poss. Volcano	y1	y2	y3	y4	y5
1668	0.070	0.35	Shikotsu				*	
1695	0.073	0.24	Komagatake ?Serua ?Hekla ?					
1732	0.152	0.91	unknown (Lanzarote ?)		*		*	
1755	0.099	0.64	Taal ? or Katla ?				*	
1762	0.065	0.26	unknown					
1783	**1.00**	**1.00**	Laki (Grimsvoetn)	*	*	*	*	*
1810	0.093	0.35	unknown tropical				*	
1815	0.071	0.65	Tambora	*		*		*
1822	0.109	0.60	Galunggung					*
1832	0.145	0.95	Babuyan Claro ?	*				
1884	0.032	0.11	Krakatau					
1900	0.093	0.66	unknown					*
1912	0.281	0.98	Katmai-Novarupta	*	*	*	*	*

TABLE 4.1 – Parallel between the detected events from our method (see Figure 4.7) and date of known volcanoes found in the literature (e.g. [Wastegard and Davies, 2009]). Note that 20th century records quite likely only show Katmai-Novarupta, while others, after 1912, are considered as spurious due to anthropogenic noise. The second column gives the relative amplitudes comparatively to the biggest event (i.e. the Laki eruption in 1783). The five last columns show whether or not the pulse like signal was detected using a univariate procedure on each time series.

6 Discussion

In this article, we introduced a state space model which allows for the extraction of timing and amplitude of pulse-like events in the presence of trends and noise. The algorithm developed here is applied to multivariate time series with a common hidden forcing. Beyond the problem of detection of the impact of volcanic eruptions on temperature time series, we believe that this type of statistical procedure is flexible enough to be able to work equally well with other time series consisting of large amplitude pulses superimposed on slowly changing trends as may be found in hydrology.

It is possible, or even desirable, to further extend this type of extraction model. In paleoclimatology, it is very rare to have perfect chronologies. Additionally, the time resolution (temporal sampling) can vary across the multitude of different types of records and time series that cover the period of interest. For instance, ice-core chronologies that were established by counting visible/detectable annual layers might be get off-track through missing or repeated layers through wind action. This potential "drift" in chro-

nologies needs to be corrected. For the purpose of a cleaner introduction of our method, we have not dealt with this issue. But before the method can be applied to the full set of ice core records, clearly, their chronologies need to be synchronized.Clustering and other techniques could make this process significantly more objective and reproducible. We are currently exploring such methods. Further, we are adapting our algorithm so that it could deal with non-regular time steps and thus could combine low-resolution (say annual) with high-resolution (full seasonal resolution) data without having to average time series ahead of time. Finally, we are re-evaluating the assumption that the unpredictable noise ϵ_j in (4.3) are no longer iid, but have a spatial structure in their covariance. These next efforts illustrate that different degrees of complexity can be added to our current model. The challenge for the geophysicist is to ensure that all information is used properly, while the challenge for the statistician is to keep the extended model simple enough for easy interpretation, and for the procedure to be able to actually estimate the necessary parameters. Overall, such collaborations between statisticians and geoscientists have the potential of advancing (in this case climate) research ([Hughes and Ammann, 2009]) by improving the quantitative treatment of the diverse records and by introducing a formal way of handling uncertainties.

7 Appendices

General remarks and notations about the estimation procedure

To identify and interpret the parameters β_j in (4.3), we have to force the variance of $x(t)$ to be equal to one. This constraint imposes that σ_v and μ_v has to obey the relationship $1-\alpha^2 = \sigma_v^2\ \pi + \mu_v^2\ \pi\ (1-\pi)$, because of (4.4) and (4.5). This implies $\sigma_v^2 = \frac{1}{\pi}(1-\alpha^2 - \mu_v^2\ \pi\ (1-\pi))$.

The matrices H and Φ and the random vectors E_t and E_t^* in equations (4.6) and (4.7) are defined in as follows

$$H = \begin{bmatrix} 0 & \beta_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \vdots & 0 & 0 & 0 & \dots & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & \beta_J & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

and

$$\Phi = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & \alpha & 0 & 0 & 0 & 0 \\ 0 & 0 & B_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & B_J \end{bmatrix}, \text{ with } B_j = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

with $E_t = (\epsilon_1, \epsilon_2, \ldots, \epsilon_J)^T$ and $E_t^* = (0, 0, E_{f_1}^T(t), \ldots, E_{f_J}^T(t))^T$. where $E_{f_j}(t)$ follows a zero-mean bivariate normal distribution with covariance $\lambda_j \sigma_j^2 \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix}$. Here H corresponds to a $J \times [2(J+1)]$-matrix, E to a J-vector, Φ to a $[2(J+1)] \times [2(J+1)]$-matrix and E^* to a $2(J+1)$-vector.

The estimation of the parameters is accomplished by sequential updating the system and observational equations in a similar manner as presented in appendices Guo et al. [1999] and Naveau et al. [2003]. However, there are slight differences since this algorithm is extended to a multivariate framework. See details below in 7.

Although the smoothing parameter can be chosen using an automatic method through cross validation techniques, λ_j remains a free choice by the user in our algorithm. From experience, we find that a small value of λ_j such as 0.01 works well with our simulated and real data sets. Using the same value of λ_j across all data sets aided in the comparison of the results.

Multivariate MKF

For any given values of the parameters of interest, the first step of this multivariate extension of the multiprocess Kalman Filter begins with an initial estimate of $\hat{X}(t-1|Y_{t-1})$ (resp. $\hat{\Sigma}(t-1|Y_{t-1})$) which are defined as the expectation of X_{t-1} conditioned on the observations $Y_{t-1} = (y_1, \ldots, y_{t-1})$ (resp. the variance of X_{t-1} conditioned on the observations Y_{t-1}). The estimation of the parameters is carried out by performing the following steps.

1. Conditioned on $Z_{t-1,i} = (Y_{t-1}, I_t = i)$, the distribution of $\hat{X}_t$ is characterized by the first two moments :

$$\begin{aligned} \hat{X}(t|Z_{t-1,i}) &= \mathbb{E}[X_t|Z_{t-1,i}] \\ &= \Phi\hat{X}(t-1|Y_{t-1}) + \mathbb{E}E[E_j(t|I_t = i)] \end{aligned}$$

and

$$\begin{aligned}\hat{\Sigma}(t|Z_{t-1,i}) &= \mathbb{V}ar[X_t|Z_{t-1,i}] \\ &= \Phi\hat{\Sigma}(t-1|Y_{t-1})\Phi^T + Cov[E_j(t|I_t=i)]\end{aligned}$$

$$\begin{aligned}\mathbb{E}[E_j(t|I_t=i)] &= (\mu_v(t),\mu_v(t),0,\ldots,0)^T, \text{ if i=1} \\ &= (0,0,0,\ldots,0)^T, \text{ else}\end{aligned}$$

$$Cov[E_j(t|I_t=i)] = \begin{bmatrix} \Sigma_v & 0 & 0 & 0 \\ 0 & Cov(E_{f_1}(t)) & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & Cov(E_{f_J}(t)) \end{bmatrix}$$

$$\Sigma_v = \begin{bmatrix} \sigma_v^2 & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 \end{bmatrix}$$

Recall that $\mathbf{E}_{f_j}(t)$ is a 2-dimensional vector with zero mean and covariance elements $\lambda_j\sigma_j^2/[(i+k-1)(i-1)!(k-1)!]$.

2. The predicted distribution of $Y_t = (y_1(t), y_2(t), \ldots, y_J(t))^T$ conditioned on $Z_{t-1,i}$ is also characterized by the first two moments :

$$\begin{aligned}\mathbb{E}[Y_t|Z_{t-1,i}] &= H\hat{X}(t|Z_{t-1,i}) \\ \mathbb{V}ar[Y_t|Z_{t-1,i}] &= H\hat{\Sigma}(t|Z_{t-1,i})H^T + \mathbb{V}ar(\epsilon),\end{aligned}$$

where

$$\mathbb{V}ar(\epsilon) = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_J^2 \end{bmatrix}.$$

The distribution of $(Y_t|Y_{t-1})$ is a mixture of multivariate normal distributions ($\mathcal{MVN}$) :

$$\begin{aligned}&(1-\pi)\mathcal{MVN}(\mathbb{E}[Y_t|Z_{t-1,0}], \mathbb{V}ar[Y_t|Z_{t-1,0}]) \\ &\quad +\pi\mathcal{MVN}(\mathbb{E}[Y_t|Z_{t-1,1}], \mathbb{V}ar[Y_t|Z_{t-1,1}]).\end{aligned}$$

3. The posterior probability of index variable $I_t = i$ conditioned on Y_t is then given by
$Pr(I_t = 0|Y_t) = \frac{(1-\pi)\mathcal{MVN}(\mathbb{E}[Y_t|Z_{t-1,0}],\mathbb{V}ar[Y_t|Z_{t-1,0}])}{Pr(Y_t|Y_{t-1})}$,
and
$Pr(I_t = 1|Y_t) = \frac{\pi\mathcal{MVN}(\mathbb{E}[Y_t|Z_{t-1,1}],\mathbb{V}ar[Y_t|Z_{t-1,1}])}{Pr(Y_t|Y_{t-1})}$.

4. Given Y_t we can update the first two moments of X_t conditioned on I_t :
$\hat{X}(t|Y_t, I_t = i) = \hat{X}(t|Z_{t-1,i}) + \hat{\Sigma}(t|Z_{t-1,i})H^T[\mathbb{V}ar(Y_t|Z_{t-1,i})]^{-1}[Y_t - Y(t|Z_{t-1,i})]$,
and
$\hat{\Sigma}(t|Z_t) = \hat{\Sigma}(t|Z_{t-1,i}) - \hat{\Sigma}(t|Z_{t-1,i})H^T\mathbb{V}ar[Y_t|Z_{t-1,i}]^{-1}H\hat{\Sigma}(t|Z_{t-1,i})$.

5. Marginal distribution of $\hat{X}(t|Y_t)$ is a mixture of multivariate normals. We approximate the mixture by a multivariate normal with the same first two moments.
$\hat{X}(t|Y_t) = (1-\pi)\hat{X}(t|Z_{t,0}) + \pi\hat{X}(t|Z_{t,1})$,
and
$\hat{\Sigma}(t|Y_t) = (1-\pi)\hat{\Sigma}(t|Z_{t,0}) + \pi\hat{\Sigma}(t|Z_{t,1}) +$
$(\hat{X}(t|Z_{t,i}) - \hat{X}(t|Y_t))(\hat{X}(t|Z_{t,i}) - \hat{X}(t|Y_t))^T$.
This steps allows us to continue sequential updating.

This work was supported by the NSF-CMG (ATM-0327936) grant, the NCAR Weather and Climate Impact Assessment Science Initiative, the EU-FP7 "AACQWA" Project (www.acqwa.ch) under contract Nr 212250, the PEPER-GIS project, the ANR-MOPERA project and the GEOmon project (www.geomon.eu). The authors would also like to credit the contributors of the R project. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

8 Validation a-posteriori de la méthode : caractéristiques des résidus

Afin de tester les sorties de notre modèle sur les différentes données, nous présentons dans cette partie, l'étude des résidus $\epsilon_j(t)$ de l'équation (4.1). L'étude de ces résidus montre l'adéquation de la méthode utilisée pour résoudre la décomposition du modèle (4.1) avec les données d'étude. Pour cela, nous testons sur ces résidus les hypothèses de gausiennité et d'indépendances émises lors de la résolution par filtre de Kalman. Nous présenterons successivement les illustrations de résidus, celle des *Quantile-Quantile plot*, qui permet, ici, de comparer la distribution des résidus avec une distribution Gausienne, et enfin la fonction d'autocorrelation, qui permet d'estimer l'indépendance temporelle des résidus (consulter [Bourbonnais and Terraza, 2004]).

8.1 Résidus des données simulées

Nous présentons tout d'abord (Figure 4.8) une nouvelle illustration de la méthode présentée dans ce chapitre. Les séries noires correspondent aux entrées du modèle, les séries de couleurs correspondent aux sorties. Les séries bleues représentent les tendances extraites des séries initiales (en noir). Les signaux rouges communs à toutes les séries représentent les signaux cachés extraits.

Les Figures 4.9, 4.10 et 4.11 présentent successivement les séries de résidus, les *QQ-plots* et les fonctions d'autocorrélation associés. Au regard des différents résultats présentés ici, les hypothèses sur les résidus, émises lors de l'élaboration du modèle (4.1) sont bien vérifiées en sortie. Les résidus présentent une distribution Gaussienne, sans dépendance temporelle.

8.2 Résidus des données de l'application

Nous présentons ici les résultats sur les résidus des données réelles de l'application aux séries de sulfate volcanique. Afin de ne pas charger l'article, ces résultats n'y avaient pas été présentés. Ici encore les hypothèses semblent avoir été respectées, l'illustration est faite via les Figures 4.12, 4.13 et 4.14. Ces Figures illustrent le fait que les hypothèses sur la caractère gaussien et l'indépendance des résidus des séries de l'application aux données de carottes de glace peuvent être considérées valables.

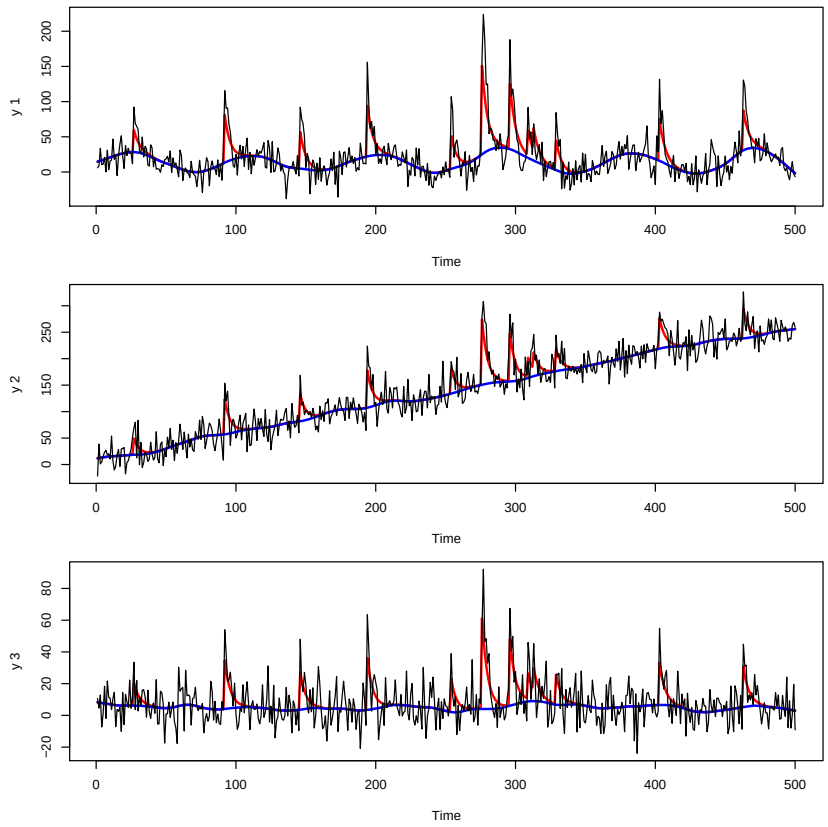

FIGURE 4.8 – Illustration sur une même figure de la méthode présentée dans ce chapitre. Les séries en bleue représentent les tendances extraites des différentes séries. Les séries en rouge correspondent au signal caché, commun à toutes les séries.

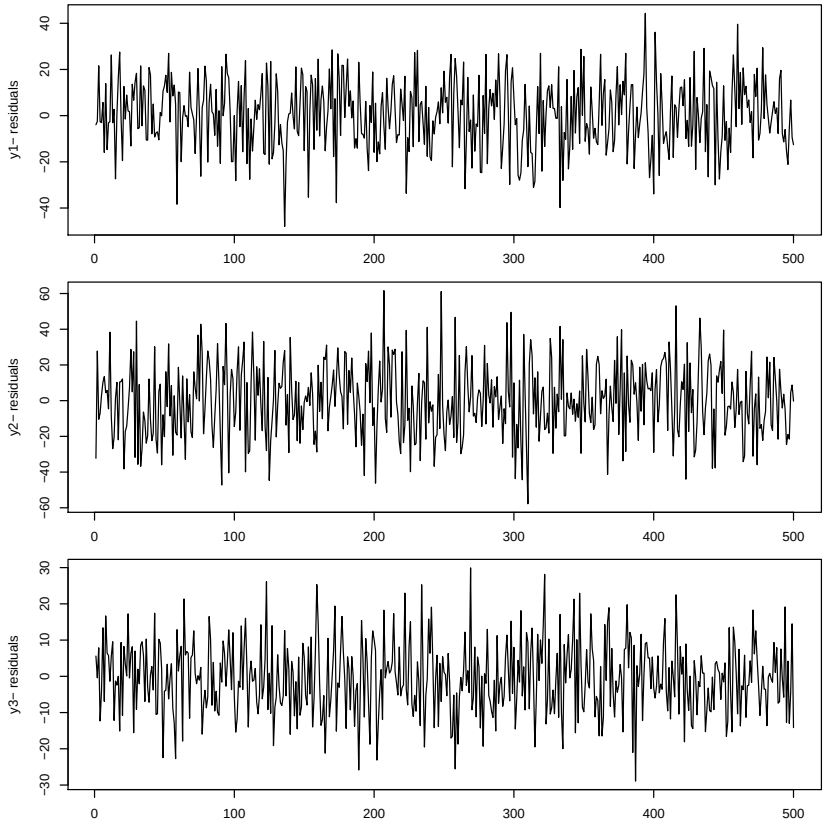

FIGURE 4.9 – Signaux $\epsilon_j(t)$ de l'équation (4.1) récupérés à partir de séries simulées.

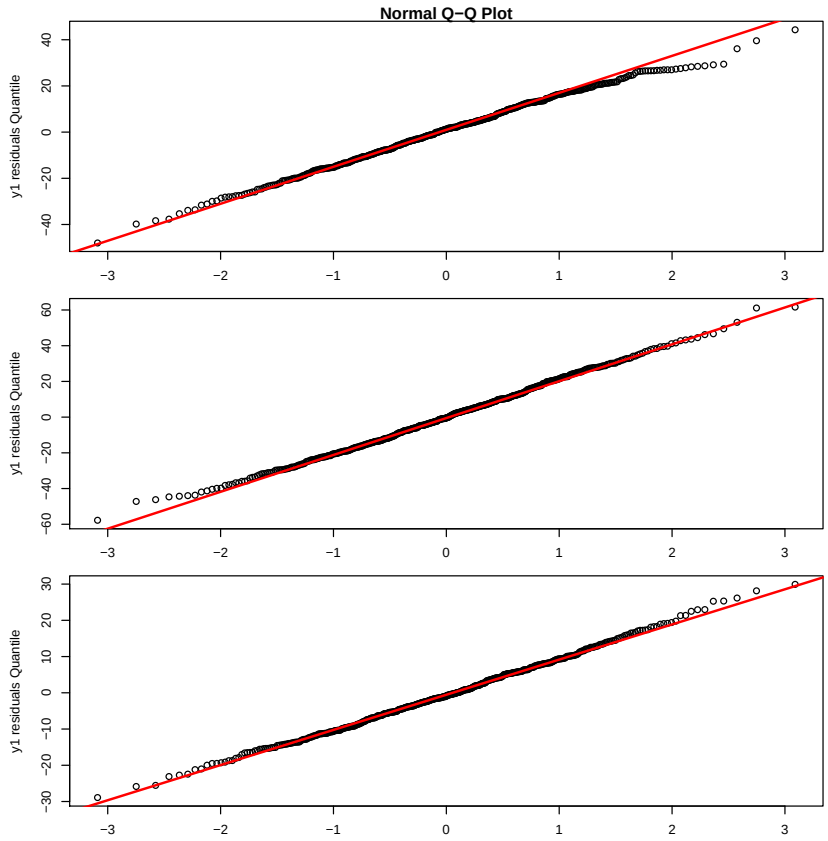

FIGURE 4.10 – *QQ-plots* des séries $\epsilon_j(t)$ des données simulées de la Figure 4.9. On remarque que la distribution de résidus (en noir) est assimilable à une distribution Gaussienne (en rouge).

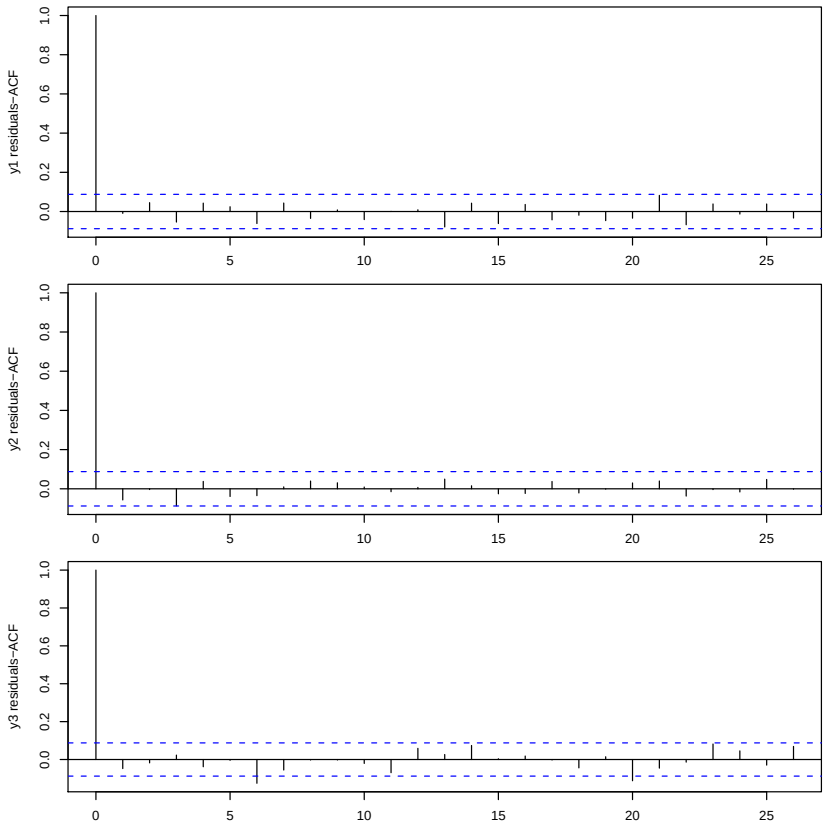

FIGURE 4.11 – Fonction d'autocorrelation des séries $\epsilon_j(t)$ des données simulées de la Figure 4.9. On peut considérer que, en dehors de l'ordre 0, les autocorrélations du signal sont nulles. Ce qui illustre, pour chaque série j considérée, le caractère indépendant des différents $\epsilon_j(t)$.

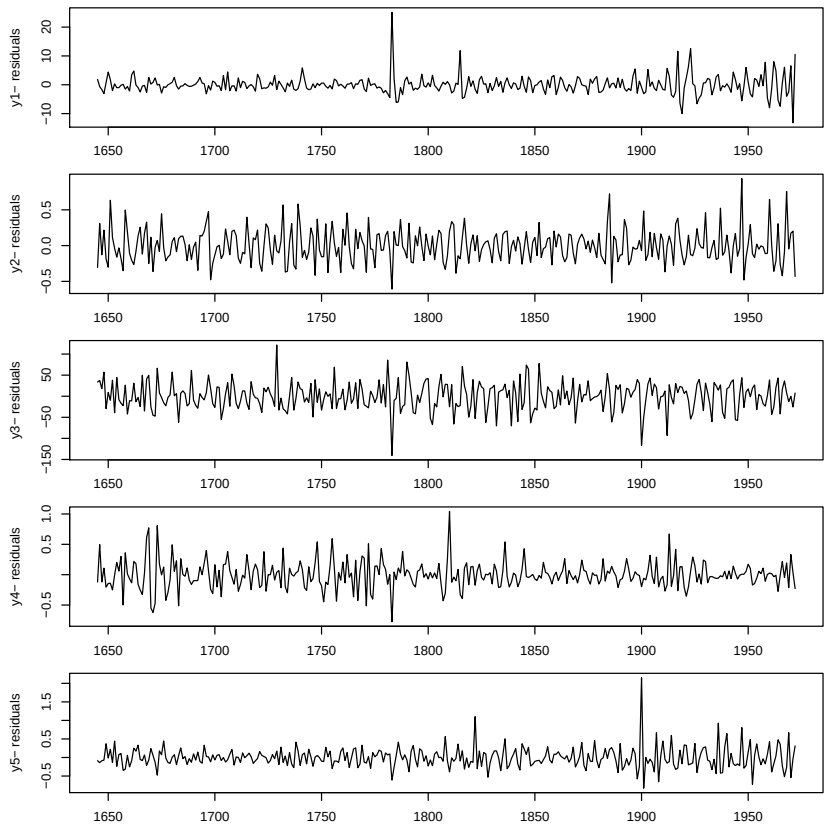

FIGURE 4.12 – Signaux des résidus issus de l'extraction sur les cinq séries de sulfate du Groenland.

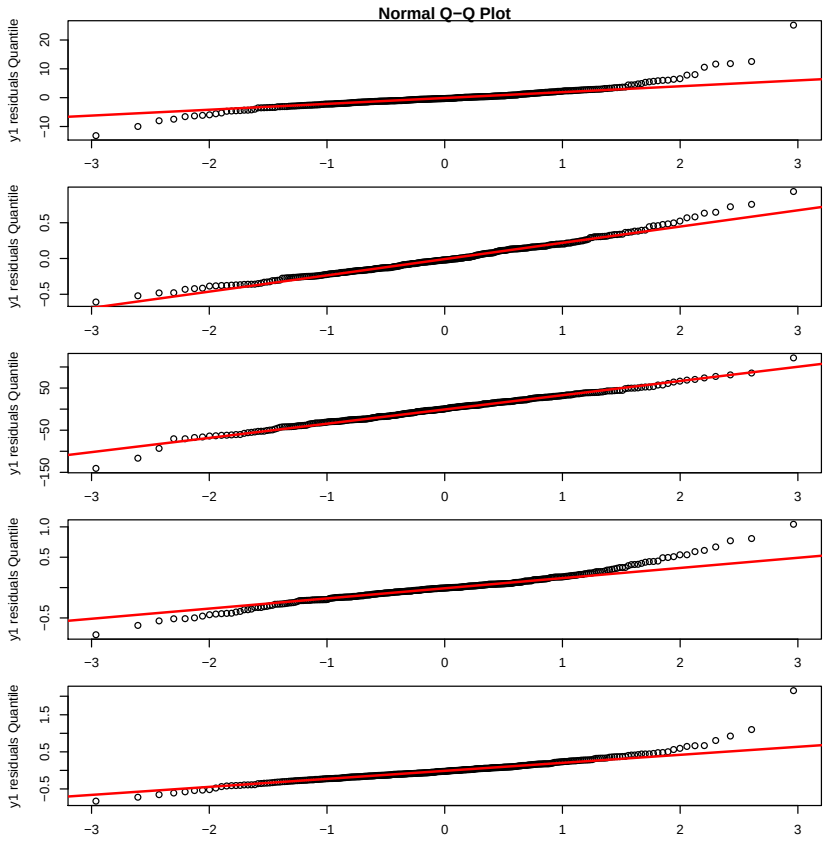

FIGURE 4.13 – Présentation des *QQ-plots* des séries de résidus extraient des séries de carottes de glace (voir Figure 4.12). La Gaussiennité des résidus est ici moins évidente. Cependant, si on néglige l'évènement du Laki en 1783 et la période récente (à partir de 1920, les mesures sont considérées moins précises), alors les distributions présentées devraient apparaitre davantage Gaussienne.

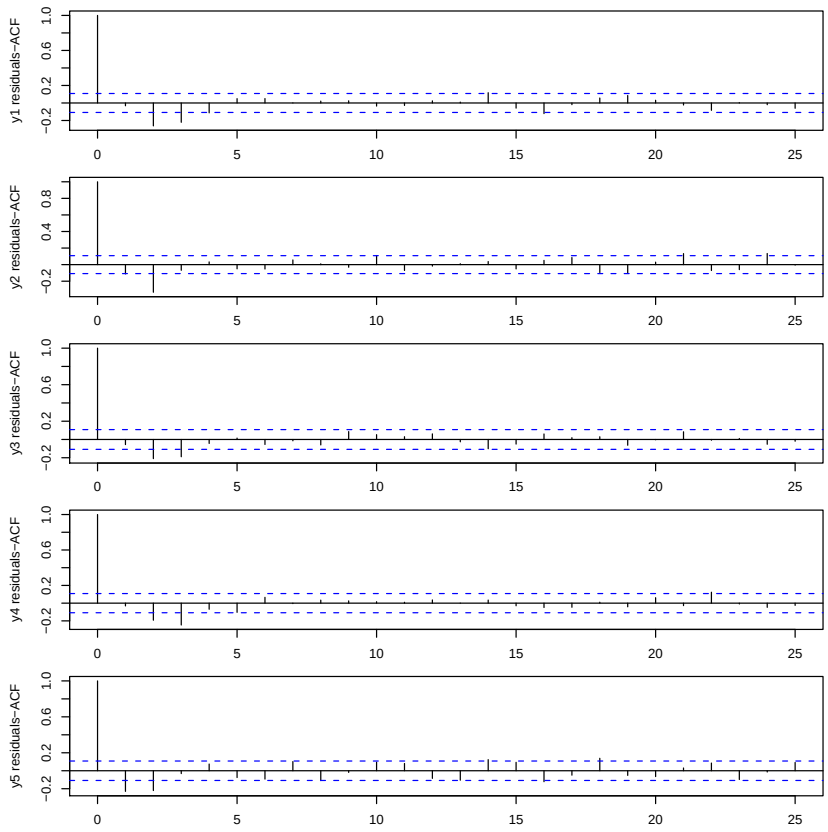

FIGURE 4.14 – Présentation de la fonction d'autocorrélation des séries de résidus extraient des séries de carottes de glace (voir Figure 4.12). Ici, l'indépendance est moins forte que sur les données simulées : on peut remarquer une très faible autocorrélation d'ordre 2, mais qui ne semble pas significative.

Chapitre 5

Détection de ruptures dans des séries multivariées non stationnaires : Application à l'Onset de la Mousson de l'Afrique de L'Ouest

Résumé : Ce dernier chapitre présente une méthode de décomposition du signal appliquée à la détection de ruptures (i.e. début de la mousson africaine dans des séries chronologiques de données climatiques). Une série j d'observation est modélisée par :

$$y_j(t) = f_j(t) + \beta_j x(t) + \epsilon_j(t), \tag{5.1}$$

où x sont des signaux de ruptures (voir Figure 5.1) d'amplitude β_j dans des séries temporelles multivariées non stationnaires présentant une tendance inconnue f_j et entachées d'un bruit ϵ. Ces ruptures interviennent à des instants aléatoires simultanés à toutes les séries et font apparaître des non linéarités dans le signal. Bien que la base du modèle probabiliste soit comparable à celle présentée dans le Chapitre 4. La détection et la caractérisation de ruptures sont différentes de celles d'évènements éruptifs et posent de nouveaux problèmes. Le signal x est modélisé par un signal $AR(1)$ dans lequel $\alpha = 1$, et donc sa variance n'est pas stationnaire ce qui amène à s'intéresser notamment à l'augmentation au fil du temps, de l'incertitude sur les détections. Nous présentons une application de cette décomposition de signal dans la détection d'un phénomène (l'onset de la mousson de l'Afrique de l'ouest) caractérisé par une rupture dans certaines variables de la dynamique atmosphérique régionale. Nous introduisons cette étude grâce à un préambule, suivi par un article accepté dans Journal of Geophysical Research.

Plan du Chapitre 2

1. **Préambule**
2. **Change-point detection in multivariate context**
3. **West African Monsoon onsets**
4. **Statistical modeling and inference**
5. **WAM results and discussion**
6. **Appendices**
7. **Validation a-posteriori de la méthode : caractéristiques des résidus**

1 Préambule

La mousson de l'Afrique de l'ouest se traduit entre juillet et septembre par un changement de direction et de régime des vents ainsi qu'une augmentation des précipitations. Ces précipitations étant la principale source d'eau de la région, l'impact y est donc très important notamment pour l'agriculture. De plus les mécanismes de déclenchement et de développement de la mousson de l'Afrique de l'ouest sont intimement liés à la naissance des cyclones tropicaux qui traversent l'atlantique (e.g.[Thorncroft and Hodges, 2001], [Maloney and Shaman, 2008]) et dévastent la région des caraîbes pendant les mois de septembre/octobre. L'onset qui est défini comme le déclenchement de mousson est un déplacement soudain des précipitations vers le nord de la région du sahel (e.g. [Sultan and Janicot, 2003]). Ces évènements marquent le début de la saison des pluies. La détection de cet instant permet de mieux comprendre les mécanismes de déclenchement de la Mousson Africaine.

L'ITD (Intertropical Discontinuity) et l'OLR (Outgoing Longwave Radiation), deux grandeurs géophysiques illustrant respectivement les vents et les précipitations, sont des marqueurs des variations annuelles de la Mousson de l'Afrique de l'ouest. Appuyés par différentes études (e.g. [Sultan and Janicot, 2003], [Janicot et al., 2008]), nous avons donc légitimement considéré que ces séries portaient la marque de l'onset, comme un instant de rupture simultanée des deux séries. Les problèmes rencontrés dans cette étude sont le fait du caractère remarquablement différent des données traitées, autant en terme de tendance que variabilité haute fréquence.

L'objectif étant la détection de l'onset à partir des séries d'ITD et d'OLR, nous avons étudié une approche multivariée prenant en considération des tendances différentes et inconnues, et des variabilités distinctes pour les deux séries. Nous faisons l'hypothèse de l'apparition à des instants aléatoires simultanés aux deux séries d'une rupture caractérisant le déclenchement de la Mousson.
Le modèle (5.1) est similaire au modèle de l'équation (4.1) à cela près que le signal x est modélisé grâce à un processus $AR(1)$ non stationnaire (voir Figure 5.1). Cette étude a permis de mettre en évidence la difficulté de détecter exactement le début de la Mousson de l'Afrique de l'ouest. Nous montrons dans la Figure 5.15 la distribution de ces détections faites sur une période de 30 ans. Il est à noter la variabilité importante des dates de l'onset.
Le modèle d'extraction mis en place permet d'estimer une tendance propre à chacune des

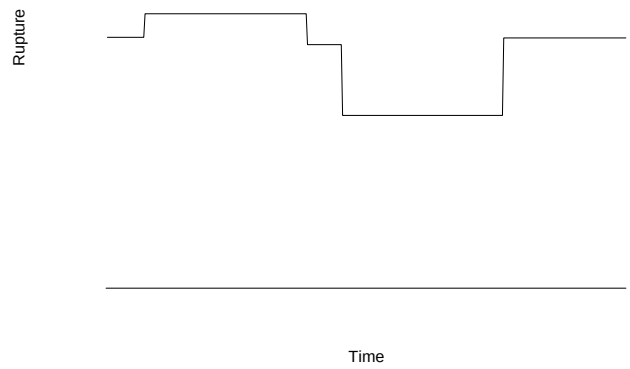

FIGURE 5.1 – Une simulation d'un signal de rupture x de l'équation (5.1) recherché dans les séries.

séries et de détecter des instants de rupture. L'application à la mousson de l'Afrique de l'ouest répond à un problème spécifique mais ne permet pas d'exploiter tout l'intérêt de ce système de décomposition de signal. En effet, les tendances extraites, propres à chacune des séries, n'ont pas de sens physique directement exploitable, seul l'extraction de la date de l'Onset est étudiée. On peut imaginer d'autres applications, telles que des applications d'homogénéisation, ou de détections de contours dans lesquelles ces tendances pourraient être étudiées tout autant que les instants de ruptures.

Dans une première section, nous revenons sur des problèmes généraux de détection de rupture dans le contexte d'une étude multivariée. La section suivante détaille le mécanisme de l'onset de la Mousson de l'Afrique de l'ouest. Enfin nous expliquons la méthode mise en place pour détecter l'instant de l'onset grâce au développement d'un filtre de Kalman dans un cadre non stationnaire et non linéaire. Nous présentons quelques résultats de validation et les comparons avec des résultats d'études précédentes sur le sujet.

Article #3 : doi :10.1029/2010JD014723 - ***Journal of Geophysical Research***

Inferring change-points and non-linear trends in multivariate times series : Application to West African Monsoon onset timings estimation

Julien Gazeaux[1], Emmanouil Flaounas[1], Philippe Naveau[2], Alexis Hannart[3] [1]UPMC Univ. Paris 06 ; Université Versailles St-Quentin ; CNRS/INSU, LATMOS-IPSL, France, Paris
[2]Laboratoire des Sciences du Climat et de l'Environnement (LSCE), IPSL, CNRS/CEA, France, Gif-Sur-Yvette
[3]Institut Franco-Argentin d'études du Climat et ses Impacts (IFAECI), Universitad de Buenos Aires,

abstract

Time series in statistical climatology are classically represented by additive models. For example, a seasonal part and a linear trend are often included as components of the sum. Less frequently, hidden elements (e.g., to represent the impact of volcanic forcing on temperatures) can be integrated. Depending on the complexity and the interactions among the different components, the statistical inference challenge can quickly become difficult, especially in a multivariate context where the timings and contributions of hidden signals are unknown. In this article we focus on the statistical problem of decomposing multivariate time series that may contain both non-linear trends and change-points (discontinuities), the change points being assumed to occur simultaneously in time for all variables in the multivariate analysis. The motivation for such a study comes from the statistical analysis of the West African Monsoon (WAM) phenomenon for which unknown pre-onset and onset dates occur each year. The impacts of such onsets can be statistically viewed as yearly change-points that affect, almost synchronously, trends in observed time series such as daily Outgoing Longwave Radiation and the Intertropical Discontinuity. Our proposed model corresponds to a multivariate additive model with non-linear trends

and possible yearly discontinuities, modeling the onsets. An inference scheme based on a non-linear Kalman filtering approach is proposed. It enables to identify the different parts hidden in the original multivariate vector. Our inference strategy is tested on simulated data and applied to the analysis of the WAM phenomenon during the period 1979-2008. Our extracted onset dates are then compared to the ones obtained from past studies.

keywords : change-point, non-stationary, non-linear statistics, Kalman Filter, West African Monsoon

2 Change-point detection in a multivariate context

Climate time series can often be affected by artificial shifts and/or natural discontinuities due to changes in measurement conditions for the former and physical changes for the latter. To detect and interpret such abrupt and local shifts, many so-called change-point statistical procedures have been developed and studied in times series analysis (e.g., Beaulieu et al. [2007]). Current methods simultaneously determine the number of change-points and infer their positions. Beyond the specific context of homogenization in climatology (e.g. Caussinus and Mestre [2004]), the change-point problem is a vast and extensively treated domain of statistics, with applications in econometrics, finance, biology, agronomy and hydrology among others. A general review of most common approaches can be found in work by Reeves et al. [2007]. In a frequentist context, Davis et al. [2006] provided a genetic optimization algorithm to extract change-points in non-stationary univariate time series using Minimum Description Length principle assuming piecewise Auto-Regressive models. Within a Bayesian framework (e.g., Chib [1998] and Lavielle and LeBarbier [2001]), Hannart and Naveau [2009] recently proposed a fast and efficient algorithm to perform a multiple change-point detection technique based in segmenting the time series into subsequences and on prior knowledge derived from past homogenization studies. One common assumption in most change-point algorithms is that smooth trends have been removed prior to applying a chosen detection procedure. Basically this means that the data under study are assumed to come from a zero-mean stationary signal affected by an unobserved change-point process that characterizes the timing and the amplitudes of the hidden shifts. For the practitioner, this assumption implies a procedure that has two independent steps : (a) the removal of trends and (b) the extraction of change-points. This makes sense in homogenization because meteorologists (e.g., Caussinus and Mestre [2004]) classically work with pairwise differences from a set of temperatures records and an artificial shift in one time series should remain in the differences while smooth trends disappear by differencing. In other applications this two-step

strategy may not be optimal and the assumption of a zero-mean stationary signal with shifts in step (b) can be challenged. To illustrate this issue one can imagine two idealized cases. First, two time series, say of daily methane and ozone recorded at the same station, have a few common artificial discontinuities, e.g. due to changes in the station location. Making the difference between these two series would not necessarily remove trends because methane and ozone may have different low frequency signatures. The second case could be of two temperature recordings over a climatic homogeneous region in which spatially coherent abrupt changes occur synchronously in time (may be, due to weather regimes modifications or network-wide changes in observing practice). Here having synchronous breakpoints implies that taking the difference between the two time series could greatly diminish the hidden shifts intensity and consequently makes it impossible to find change-points from this difference. For these two examples one option could be to pre-process each series independently in order to remove the low frequency components. Subtracting these low frequency components in order to work with zero-mean stationary signals can still be an issue because large change-points induce a strong bias in the overall background variance estimation and consequently this may lead to estimation errors of these low frequency components. In addition, any estimation errors produced during the first step (the removal of trends) can propagate other estimation errors into the second step (the change-point extraction procedure). Finally a joint statistical analysis should improve the detection because the hidden signal is supposed to affect all time series (with different degree). Ideally it would be of interest to propose a global model and a general inference approach that bypasses the two-step estimation procedure. In its most general form this objective is overly complex because each time series can have its own non-linear trend and shares hidden change-points. Consequently, additional assumptions are needed and they should be driven by the application at hand.

The statistical model presented in this study is applied on both simulated and real climatological data. Section 3 provides the background theory on the real data application, which corresponds to the detection of the West African Monsoon (WAM) onset and explains the statistical problem concerning the estimation of unknown yearly onsets timings. Section 4 corresponds to the main statistical part of this work. Our statistical model is defined there and the inference scheme used to estimate unknown quantities is proposed and tested on simulated data. Then this scheme is applied on two variables representative of the WAM onset, for the period 1979-2008. The extracted onsets are compared to past results. Conclusions and perspectives are discussed in Section 5. An appendix at the end of the paper provides the technical parts of our algorithm and technical details about our data sets.

3 West African monsoon onsets

The West African Monsoon (WAM) regulates the rainfall season and is of paramount importance for food security and local economy. The northern WAM propagation interacts with other regional climatic features (such as the African Easterly Waves) which may result to the cyclo-genesis budding within the West African coast and eventually the initiation of tropical cyclones Thorncroft and Hodges [2001].

The rain band associated to the WAM makes part of the seasonal cycle of the Inter Tropical Convergence Zone (ITCZ). Following the ITCZ intra-seasonal cycle, the WAM blows over west Africa from early Spring to early autumn, advecting humidity and regulating the overland ITCZ location. The WAM onset corresponds to the abrupt displacement of the ITCZ and the WAM towards the north and signalises the initialisation of the rainy season for the Sahel. To illustrate the relation between the WAM and the ITCZ, Figure 5.2 presents monthly averages of Outgoing Longwave Radiation (OLR) superimposed over 925 hPa wind circulation patterns for the period 1979-2008. The OLR values are taken from the National Oceanic and Atmospheric Administration (NOAA) archive (Liebmann and Smith [1996])and is used as a proxy for deep convection since low OLR values are associated to the cold cloud tops of convective systems. The OLR dataset is interpolated to a 2.5^0x2.5^0 grid and corresponds to mean daily values. The wind data are taken from the National Center for Environmental Prediction (NCEP) 2 reanalysis Kanamitsu et al. [2002], also corresponding to mean daily values interpolated to a 2.5^0x2.5^0 grid. For all months, the ITCZ is marked by low OLR values and the WAM is represented by the southwest flow. The WAM due to its charge in humidity is cooler and more humid than the northeast dry and warm Harmattan wind which originates from the Sahara desert. Hence, a zone with frontal characteristics is created which propagates according to the WAM inland penetration. This frontal zone is referred to as the Inter-Tropical Discontinuity (ITD). Due to the different direction of these two winds, the location of the ITD is determined by the zero isotach of the zonal wind. From May until early June the ITCZ is strong and located over the Guinean coast (approximately at 5^0N). Similarly, the WAM presents a weak inland intrusion and hence the ITD is located along 15N (pre-onset period). On the other hand, from July to August the ITCZ is installed over the Sahel (along 10N) and the ITD reaches 20N (post-onset period). The transition from the pre-onset period to the post-onset period is characterized by the significant weakening of convection over the entire region and is detected to occur during late June.

Taking advantage of the zonal symmetry of the ITCZ and the ITD over West Africa, Figure 5.3 shows Hovmoller diagrams of three random years (1992, 1998, 2005) for the

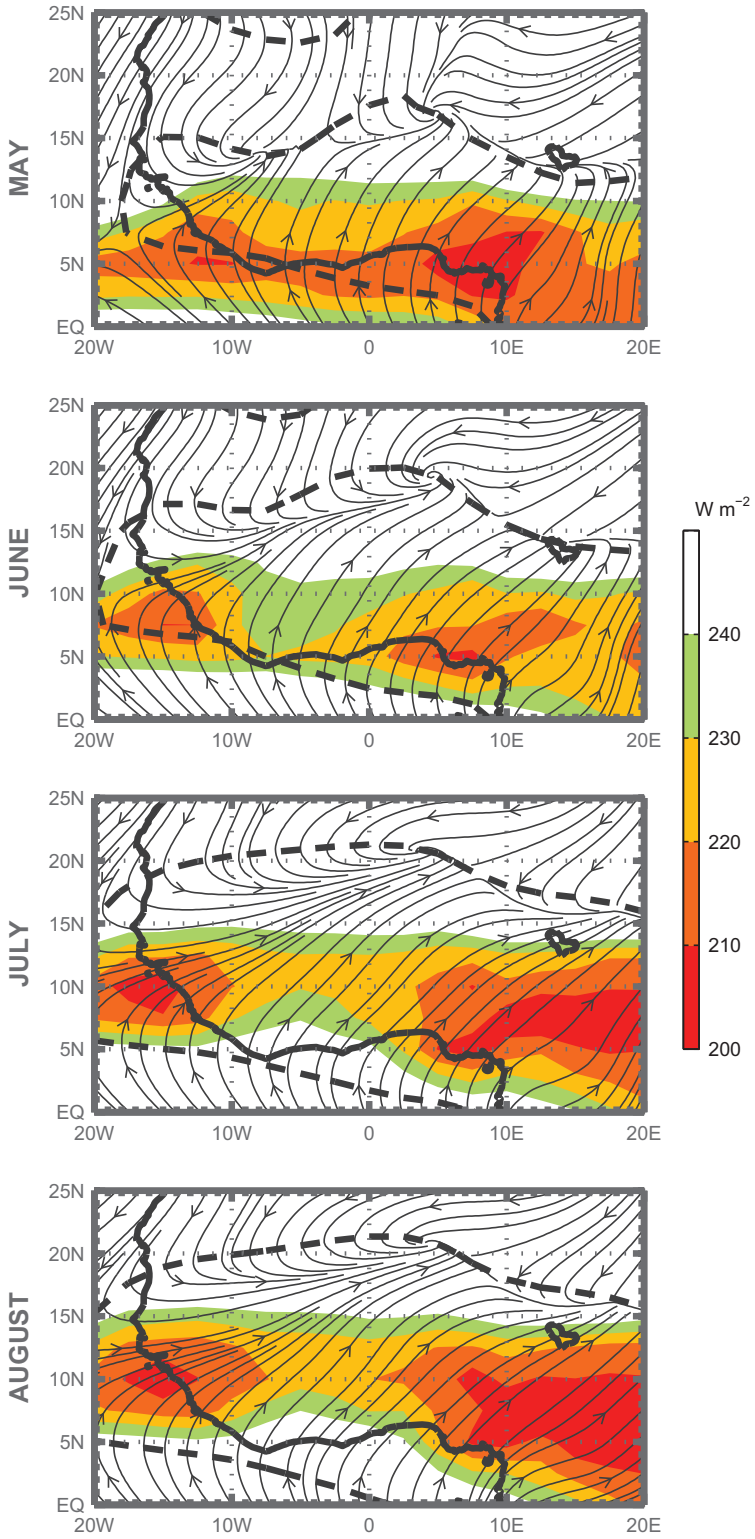

FIGURE 5.2 – Illustration of the onset phenomena : Monthly averages of the 925 hPa atmospheric circulation and OLR fields from May to August. Thick contour represents the zero zonal wind isotach

OLR values superimposing the ITD location. In these diagrams the OLR values and the ITD location are averaged between 10W and 10E and then smoothed by a moving average of 2 days to eliminate intense variability. For the three years plotted, it is important to underline the repeat of the same time-latitude pattern, for both OLR and ITD. The northward displacement of the ITD is also accompanied by the decrease of OLR values.

Previous studies of rainfall climatology Nicholson [1981], Sultan and Janicot [2000], Le Barbé et al. [2002] have put into light the intra-seasonal cycle of the ITCZ. Detecting the initialization of the rainy period over the Sahel (10N to 20N), which is characterized by abrupt changes in the regional atmospheric circulation and rainfall, is currently an object of active research Fontaine et al. [2008].

By plotting the average precipitation between 10W and 10E along 15N from the NCEP2 reanalysis database, Sultan and Janicot [2003] identified two breaks in the positive rainfall slope and they interpreted them as pre-onset dates (when the ITD reaches 15N) and onset dates (installation of the ITCZ along 10N). In Fontaine and Louvet [2006], Fontaines and Louvet analyzed rainfall data to define two precipitation indexes. The first one was based on averaging precipitation over the region (10W to 10E and the Equator to 7.5N) and the second one over the same longitude band but with different latitudes, from the Equator to 20N. Whenever the difference between these two indexes became positive for at least twenty days, an onset was considered to have taken place during the first instant of this period. Finally, Fontaine et al. [2008] studied OLR data to determine onset dates by calculating percentages of deep convection occurrences.

Inspired by these different studies, we aim at proposing a unifying statistical approach that can view such onsets as yearly change-points that affect, almost synchronously, multivariate time series. Following the aforementioned authors, we construct two time series from two databases. First of all, we construct a time series of daily OLR fields (taken from the NOAA archive) within the Sahel region (10W to 10E and 12.5N to 20N) for each year from 1979 to 2008. The 12.5N boundary of the chosen domain is justified from the fact that is far from the Guinean coast and it is strongly affected by convection over the Sahel. Hence, if before the WAM onset convective activity takes place over the Sahel it is more unlike to noise the constructed OLR time series and the statistical model to attribute a false WAM onset. Secondly, the NCEP2 reanalysis is used in order to detect the northern reach of the WAM. For this reason we calculate the mean daily location of the ITD, which corresponds to the mean latitude of the zero zonal wind isotach between 10W and 10E.

To illustrate the yearly behaviour of such data, the top panel of Figure 5.4 displays the daily time evolution of the ITD location (dark line in $latitude$) and OLR (gray line in $W.m^2$) time series from January 1992 to November 1992. From this panel, it is clear

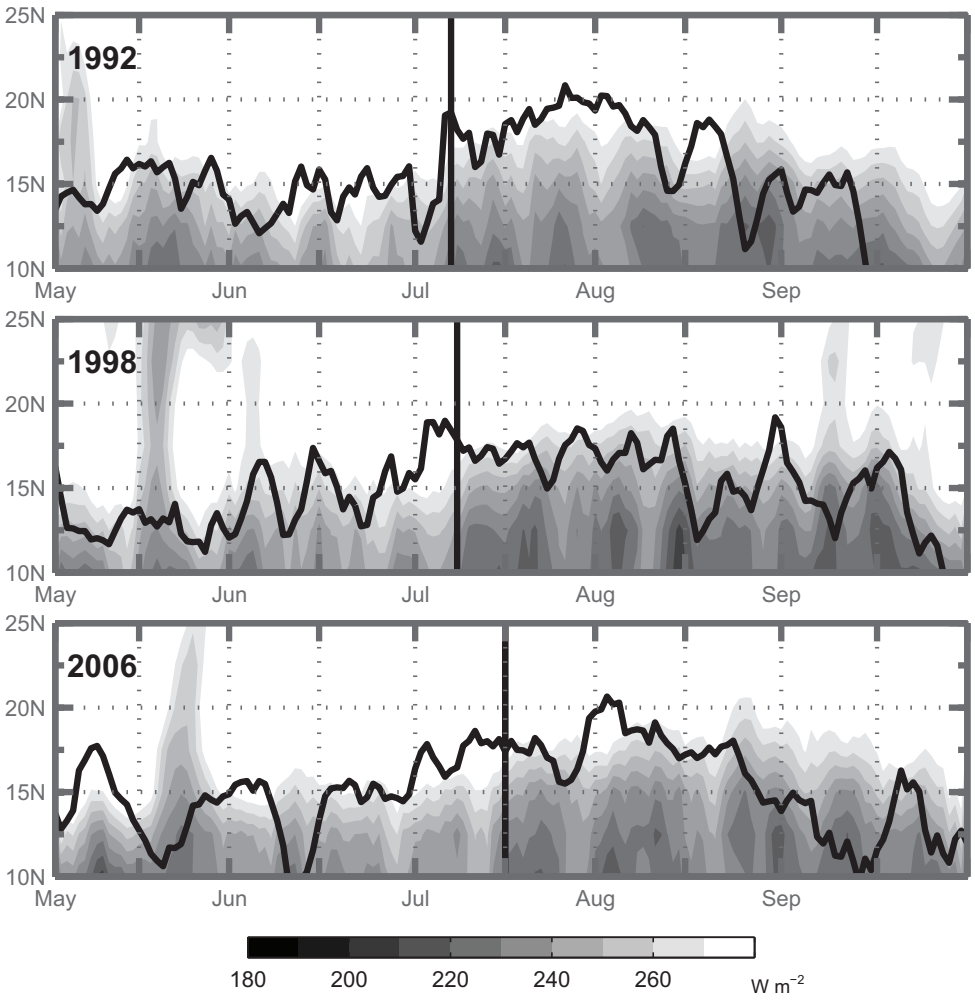

FIGURE 5.3 – Hovmoeller diagram of OLR. OLR values were averaged from 10W to 10E and smoothed by a moving average of +/- 2 days. Thick black line corresponds to the ITD position as the zero zonal wind isotach at 925 hPa. The vertical black bars represent the dates of the onset we estimated. We zoomed the time axis to better show the phenomena.

that the ITD location steadily increases until the month of August and follows by a slow decline in Autumn.

The OLR has the opposite low frequency behaviour with a stronger variability. Following the work of Fontaine and Louvet [2006] and Sultan and Janicot [2003], we postulate that these ITD and OLR signals could contain hidden change-points corresponding to the pre-onset (installation of the ITCZ along 5N) and the onset (installation of the ITCZ along 10N) dates. Hence, besides the high variability observed in Figure 5.4 and the presence of missing data, the statistical issue at hand in this paper is how can smooth behaviours as well as hidden shifts can be estimated by jointly modeling these OLR and ITD time series for the year 1992. The same question can be asked for each year in 1979-2008. To show the year-to-year variability, the years 1990, 1992, 1998 and 2006 are plotted in the different panels of Figure 5.4, respectively. It is an understatement to say change-points and trends are not easily identifiable by visual inspection of Figure 5.4 and non-trivial statistical analysis are needed.

4 Statistical modeling and inference

Our statistical model takes its roots in the classical family of state-space models. This means that a two-layer structure provides the modeling foundation. The first layer corresponds to the data while the second layer represents the processes of interest which live in the so-called state-space. The first and second layers are observed and hidden, respectively. A large body of work on inverse problems, data assimilation and Bayesian modeling is based on this idea of state-space modeling. For example, the well-known Kalman Filter (KF) allows to estimate the hidden state of a dynamical linear system (e.g. [Kalman, 1960], [Welch and Bishop, 1995], and [Meinhold and Singpurwalla, 1983]). The KF has been extended in many ways to take into account non linearities and to deal with large data sets. For example, the book of Evensen [2006] treats the Ensemble Kalman Filter for data assimilation.

From a methodological point of view, our proposed statistical model stems from the work of Guo et al. [1998] who studied an extracting procedure, not for change-points but for pulse-like signals in univariate hormone time series. The generic shape of the hidden signal in Guo et al. [1998] corresponded to a peak followed by a sharp decrease while a step-wise function is the object of interest in most change-point procedures. In Gazeaux et al. [2009], we improved Guo's approach by extending it from the univariate case to the multivariate case and by applying it to the problem of volcanic forcing extraction from multivariate proxy data. Now, by building on the multivariate approach studied by Ga-

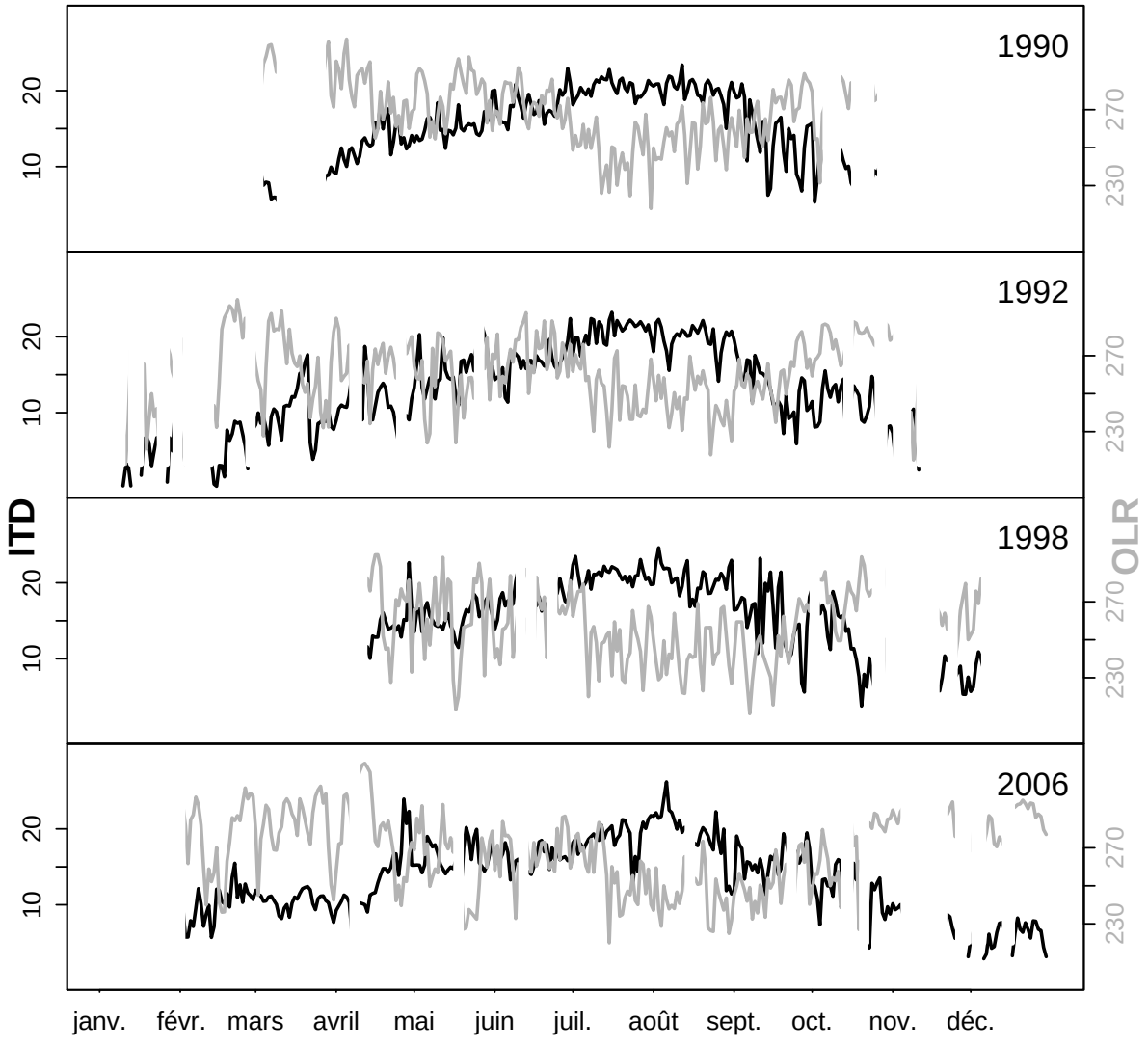

FIGURE 5.4 – Daily Outgoing Longwave Radiation (OLR) and Intertropical Discontinuity (ITD) time series for four different years 1990 (top panel), 1992 (second panel), 1998 (third panel) and 2006 (bottom panel). The dark and grey lines correspond to ITD and OLR data, respectively. The missing values in ITD are due to the difficulty to calculate the latitude of the zero zonal wind. The ITD unit is $latitude$ whereas OLR is $W.m^2$

zeaux et al. [2009] we propose to capture change-points and smooth trends. Due to the altered nature of the extracted signal and the model constraints imposed by our Monsoon application, this extension is far from trivial. A new model is needed and the statistical inference procedure has to be modified. Our proposed model corresponds to a multivariate additive model with non-linear trends and possible yearly discontinuities, the latter captured yearly onsets. While an auto-regressive cubic spline representation is used to depict different smooth trends, another auto-regressive model with non-continuous innovations mimics the change-points dynamic. Blending together these two auto-regressive models offers a modeling flexibility and removes some classical hypotheses, no linear assumption is required. To balance this flexibility in the low frequency part of the spectrum, we impose that the unknown breakpoints occur synchronously in time in all variables, see the WAM onsets application.

Concerning our notations, $y_j(t)$ represents the value of the jth variable of interest for day t. For example, $y_1(t)$ and $y_2(t)$ could correspond to the daily OLR and ITD values in 1990, see the top panel of Figure 5.4. Such random variables are assumed to come from the following additive model :

$$\begin{aligned} y_j(t) &= f_j(t) + \beta_j x_t + \epsilon_j(t), \\ &\text{with } j = 1, \ldots, J \text{ and } t = 1, \ldots, T,, \end{aligned} \tag{5.2}$$

where $f_j(t)$ represents the smooth trend specific to the jth time series, x_t the change-points signal common to all time series, β_j the scaling factor of the x_t impact to the jth time series and finally $\epsilon_j(t)$ a zero-mean independent and identically distributed (iid) Gaussian noise with variance σ_j^2. The elements of the sum (5.2) are assumed to be mutually independent. Equation (5.2) clearly indicates that each times series can have a different trend with its own noise and a common element x_t whose impact is modulated by β_j. A strong assumption of the method is, through β_j of Equation (5.2), the proportionality of the break points occurring at the same time. If all β_j have the same sign, the break have the same effect, either “positive" or “negative", on the other hand, if the β_j have opposite signs, the breaks will have opposite effects, one will be “positive" while the other will be “negative" and vice versa. We suppose in (5.2) that there is not missing value, i.e. with the constant sampling rate $t = 1, \ldots, T$. But Figure 5.4 shows missing values. Our model and our inference can handle this case by transforming the time axis $1, \ldots, T$ into $t_1, t_2, \ldots, t_k$. For sake of clarity, we still prefer to present our method with $t = 1, \ldots, T$. As the hidden signal x_t should capture the onsets dynamic, we follow the classical view

of modeling change-points as a random step-wise function. Here this step-wise behaviour is represented by an auto-regressive model of order one

$$x_t = x_{t-1} + v_t, \tag{5.3}$$

where the random variable v_t either equals zero or a zero-mean random Gaussian vector z_t with variance σ_v^2, i.e.

$$v_t = \begin{cases} 0 & \text{if } b_t = 0 \text{ with probability } 1-\pi, \\ z_t & \text{if } b_t = 1 \text{ with probability } \pi, \end{cases} \tag{5.4}$$

with $x(0)$ is set to zero, b_t is a Bernoulli iid process, either equal to one or zero with probability π and $1-\pi$, respectively. The process b_t drives the occurrences of the impulses. The Gaussian variables z_t are iid and independent of b_t.
To understand Equation (5.4), we refer to Figure 5.5. The bottom panel shows one random realization of the step-wise behaviour of x_t defined (5.3). The elements of this auto-regressive process, i.e. b_t and v_t, are displayed in the top and middle panels. Although auto-regressive, the process x_t is zero-mean but not stationary because its variance increases linearly with time, $\mathbb{V}\text{ar}(x_t) = \pi\, t\, \sigma_v^2$. In our WAM application, this is not a fundamental issue because the yearly probability of observing a change-point π is very small, we expect to have one or two change points (pre-onset and onset) per year, this implies that the yearly largest $\mathbb{V}\text{ar}(x_t)$ should be about $2\sigma_v^2$ and does not explode with time. This also justifies that we analyze our data year-per-year and not the entire period 1979-2008 in one run (this is compounded with the fact that yearly trends have a strong year-to-year variability, see Figure 5.4). Hence the hidden step-wise x_t obtained from Equation (5.3) is unlikely to produce its own trend. This implies that only the component $f_j(t)$ in Equation (5.2) should capture the low frequency in $y_j(t)$.

To model the trend $f_j(t)$, we opt for a cubic smoothing spline representation (Wahba [1978]). The latter can be described as a multivariate auto-regressive model of order one (Wecker and Ansley [1983])

$$\mathbf{F}_j(t) = B\mathbf{F}_j(t-1) + \mathbf{E}_j(t), \tag{5.5}$$

where $\mathbf{F}_j(t) = \begin{bmatrix} f_j(t) \\ f_j'(t) \end{bmatrix}$ represents a bivariate vector that includes (f_j) and its first deriva-

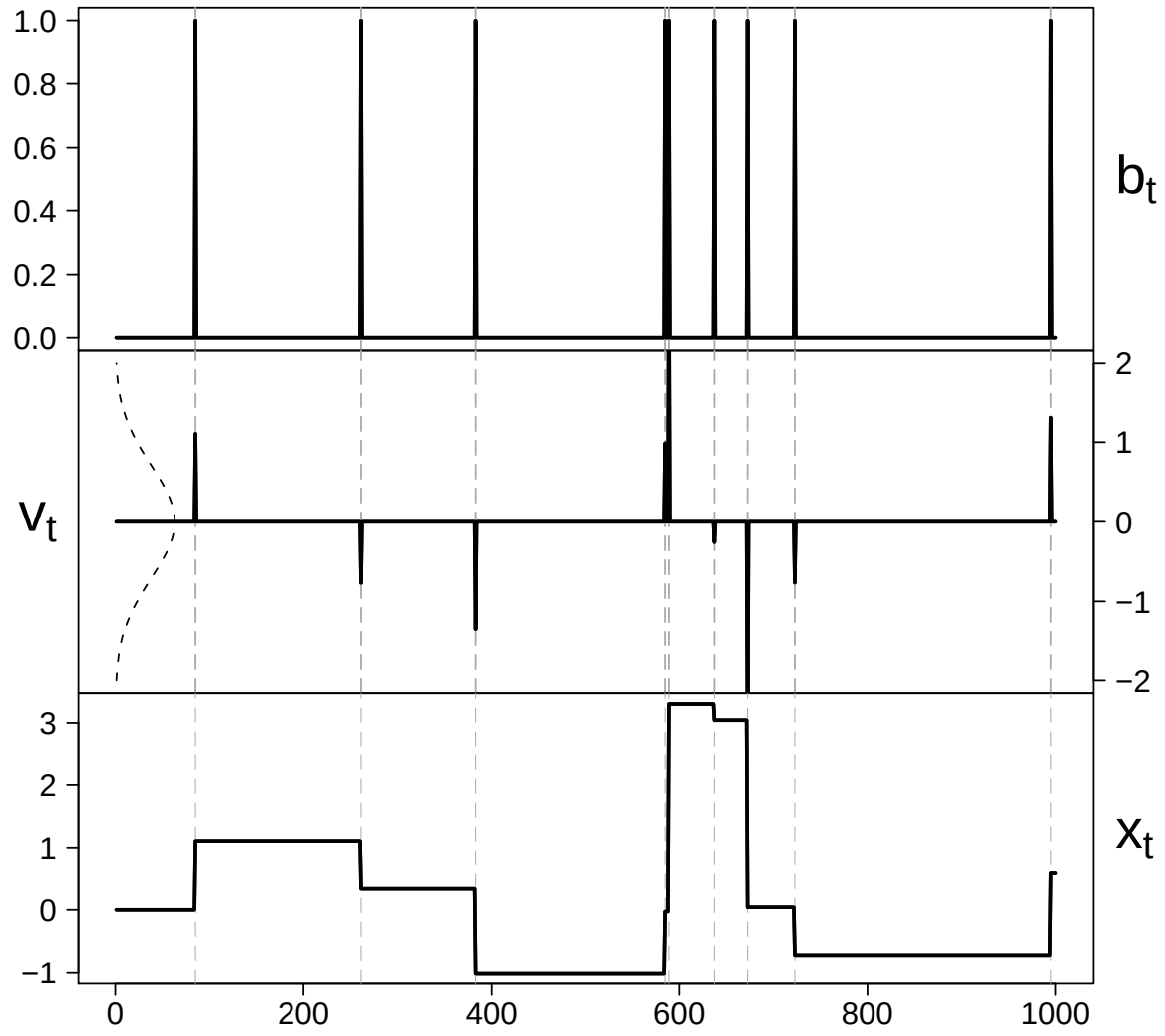

FIGURE 5.5 – Random realizations from equations (5.3) and (5.4). The top, middle and bottom panels show the Bernoulli signal b_t, the hidden impulse v_t in (5.4) and the hidden step-wise x_t obtained from (5.3), respectively.

tive (f'_j), the matrix B equals $\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}$, and $\mathbf{E}_j = \begin{bmatrix}E_{f_j}\\ E_{f'_j}\end{bmatrix}$ is a two-dimension zero-mean Gaussian vector with covariance matrix equal to $\lambda_j \sigma^2_{f_j}/(j+k-1)(j-1)!(k-1)!$ where λ_j represents the smoothing parameter and σ_{f_j} a positive constant. Equation (5.5) implicitly means that the trend f_j and its first derivative f'_j are assumed to be continuous.

Choosing equations (5.3) and (5.5) to model the unknown trends and the hidden change-points dynamic brings an important inferential benefit because our model can be rewritten as a classical linear state-space model of the form :

$$\begin{aligned} Y_t &= HX_t + E_t, \\ X_t &= \Phi X_{t-1} + E^*_t, \end{aligned} \tag{5.6}$$

where the first equality corresponds to the so-called observation equation and the second one to the so-called state equation (e.g. Meinhold and Singpurwalla [1983]). To clarify the link between equations (5.6) and (5.3)-(5.5), we write below the form of the elements of Equation (5.6) for $J = 2$, i.e. the case of the daily OLR and ITD random variables, in function of the components of equations (5.3)-(5.5)

$$\begin{aligned} Y_t = [y_1(t), y_2(t)]^T,\ X_t = [v_t, x_t, \mathbf{F}_1, \mathbf{F}_2]^T \\ \text{with } \mathbf{F}_j(t) = [f_j(t), f'_j(t)]^T \end{aligned} \tag{5.7}$$

and

$$H = \begin{bmatrix}0 & \beta_1 & 1 & 0 & 0 & 0\\ 0 & \beta_2 & 0 & 0 & 1 & 0\end{bmatrix}\ E_t = \begin{bmatrix}\epsilon_1\\ \epsilon_2\end{bmatrix},\ \Phi = \begin{bmatrix}0 & 0 & 0 & 0\\ 1 & 1 & 0 & 0\\ 0 & 0 & B & 0\\ 0 & 0 & 0 & B\end{bmatrix},$$

$$E^*_t = [0, 0, E_{f_1}, E_{f'_1}, E_{f_2}, E_{f'_2}]^T$$

,

$$\text{and } \mathbb{C}ov(\mathbf{E}_j) = \lambda_j \sigma^2_j \begin{bmatrix}1 & 1/2\\ 1/2 & 1/3\end{bmatrix}.$$

The advantage of transforming equations (5.3)-(5.5) into the state-space form Equation

(5.6) is that developments about KF can serve as building blocks for our inference procedure. By construction, the observation noise E_t and the state equation noise E_t^* are uncorrelated.

As in any KF algorithm, our main goal is to estimate the hidden state X_t given current and past observations, i.e. given the vector $Y_{1:t} = (Y_1, \ldots, Y_t)^T$. We cannot directly apply the classical KF to reach this aim because the binary vector b_t in the definition of v_t makes the hidden vector X_t non-Gaussian, see Figure 5.5. Still, as proposed in Guo et al. [1998], two ideas can be followed to remove this inference block. First, by conditioning on the value of b_t, either one or zero, we can sequentially compute the conditional expectation and variance of the two random variables $[X_t|Y_{1:t}, b_t = 1]$ and $[X_t|Y_{1:t}, b_t = 0]$ at time t, whenever those mean and variance are available at time $t-1$. The random variable $[X_t|Y_{1:t}, b_t = 1]$ corresponds to the hidden state given observations up to time t and having observed a change-point at time t and $[X_t|Y_{1:t}, b_t = 0]$ is the same except that no change-point has been observed at time t.

The second idea is to approximate the non-Gaussian distribution of $[X_t|Y_{1:t}]$, because of b_t, by a Gaussian one whose first and second moments equal of those of $[X_t|Y_{1:t}]$. This approximation that works well in practice (see the next section) allows us to update the mean and variance of the variable of interest $[X_t|Y_{1:t}]$ as follows (with obvious notations described in Appendix 6.1)

$$\begin{aligned}
\hat{X}(t|Y_{1:t}) &= q_t^0 \hat{X}(t|Y_{1:t}, b_t = 0) + q_t^1 \hat{X}(t|Y_{1:t}, b_t = 1), \\
\hat{\Sigma}(t|Y_{1:t}) &= q_t^0 \hat{\Sigma}(t|Y_{1:t}, b_t = 0) + q_t^1 \hat{\Sigma}(t|Y_{1:t}, b_t = 1) \\
&\quad + \sum_{i=0}^{1} q_t^i (\hat{X}(t|Y_{1:t}, b_t = i) - \hat{X}(t|Y_{1:t}))^2,
\end{aligned} \tag{5.8}$$

where q_t^1 (resp.q_t^0)is the occurrence probability of having (resp. not having) a breakpoint at time t given $Y_{1:t}$ and it equals (via Bayes' theorem)

$$\begin{aligned}
q_t^0 &\doteq Pr(b_t = 0|Y_{1:t}) = \frac{1-\pi}{Pr(Y_t|Y_{1:t-1})} Pr(Y_t|Y_{1:t-1}, b_t = 0), \\
q_t^1 &\doteq Pr(b_t = 1|Y_{1:t}) = \frac{\pi}{Pr(Y_t|Y_{1:t-1})} Pr(Y_t|Y_{1:t-1}, b_t = 1),
\end{aligned} \tag{5.9}$$

where $\forall i = 0, 1,\ Pr(Y_t|Y_{1:t-1})$ and $Pr(Y_t|Y_{1:t-1}, b_t = i)$ represent the conditional density of the random variables $[Y_t|Y_{1:t-1}]$ and $[Y_t|Y_{1:t-1}, b_t = i]$, respectively. As for the classical KF, the conditional mean and variance $\hat{X}(t|Y_{1:t}, b_t = i)$ and $\hat{\Sigma}(t|Y_{1:t}, b_t = i)$ can be

expressed in terms of previous expressions obtained at time $t-1$. See the Appendix for more details. The estimation of the state vector at every time $t = 1, ..T$ regarding the available observation $Y_{1:T}$ is obtained via the Fixed Interval Smoother, which is

$$\begin{aligned}\hat{X}(t|Y_{1:T}) &= \hat{X}(t|Y_{1:t}) + C_t[\hat{X}(t+1|Y_{1:T}) - \hat{X}(t+1|Y_{1:t})],\\ \hat{\Sigma}(t|Y_{1:T}) &= \hat{\Sigma}(t|Y_{1:t}) + C_t[\hat{\Sigma}(t+1|Y_{1:T}) - \hat{\Sigma}(t+1|Y_{1:t})]C_t',\end{aligned} \tag{5.10}$$

where

$$C_t = \hat{\Sigma}(t|Y_{1:t})\Phi\hat{\Sigma}(t+1|Y_{1:t})^{-1}. \tag{5.11}$$

For more details about these calculations see appendix 6.1. So far, we have assumed that the parameters $(\beta_j, \pi, \sigma_v, \sigma_{f_j})$ were known. This is not true in practice. They are derived through an iterative maximum likelihood estimation computed after a rough estimation of the trend of each time series. This method of approach is successfully used in Guo et al. [1998].

To assess the quality of our algorithm, we apply it to simulated trivariate time series defined as follows. The first, second and third smooth trends are equal to $f_1(t) = 10 + 15\sin(\pi(t+20)/90)$, $f_2(t) = 0$ and $f_3(t) = 2t$, respectively. In Figure 5.6, we can observe the three hidden trends (blue solid lines) and three random realizations (black lines) affected by zero-mean Gaussian noises with variances $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 5$ and a random change-point process with $\pi = 0.01$ and σ_v^2=1.0. The scaling coefficients are chosen such that $(\beta_1, \beta_2, \beta_3)^T = (20, 15, 20)^T$. Additional tests (available under request) show that our multivariate method represents an improvement over its univariate counterpart, i.e. applying independently our model to each individual time series.

The red lines correspond to the estimated trends with their 5- and 95-quantiles in dotted green lines.

In Figure 5.7, the three black lines represent the input of the model, whereas the different coloured lines are the output of the extraction, i.e. the estimated parts of X_t.. The top panel displays the estimated probabilities of observing change-points and the bottom panel compares the true x_t and its estimate. Graphically the timing and amplitudes of the change-point appear to be well estimated. Only the smallest shifts at approximately $t = 160$ and $t = 380$ are associated with low probabilities of about 0.2 and less than 0.1. The single simulation analysis shown in Figures 5.6 and 5.7 is obviously not sufficient to conclude about the overall performance. Rather it has to be understood as a graphical

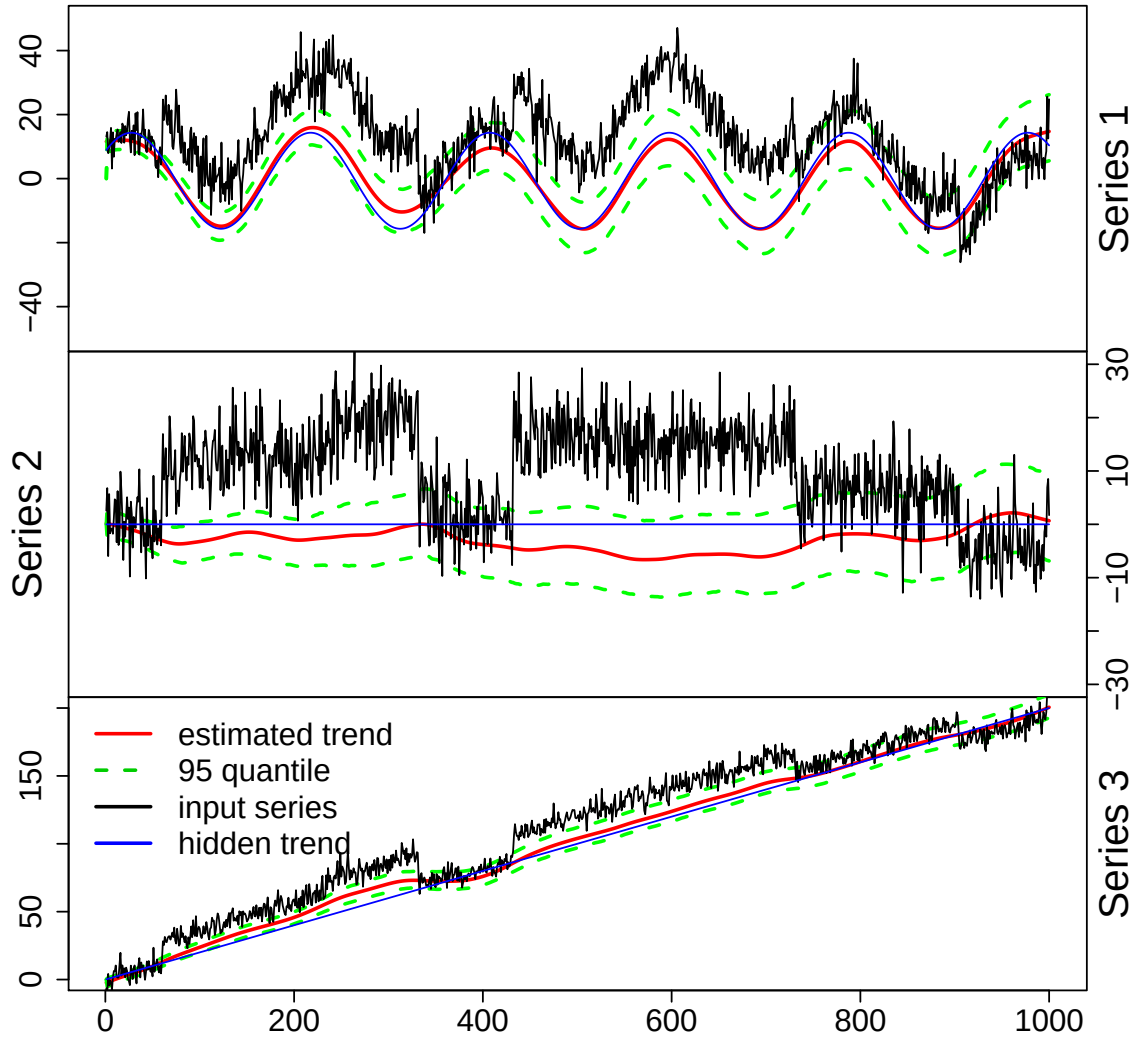

FIGURE 5.6 – Extraction obtained from three simulated time series. The blue and red lines correspond to the true and estimated trends, respectively. The 95% confidence interval is represented to the green dotted lines.

example of the possible outputs available from our approach. To improve our understanding of the limits and advantages of our method, we apply our algorithm to different sets of parameters. For each set, 500 simulations are randomly generated and boxplots of the parameters of interest are plotted. For example, when fixing $\beta = (20, 15, 20)^T$ and $\pi = 0.01$, the x-axis of Figure 5.8 corresponds to five different combinations of the triplet $(\sigma_1, \sigma_2, \sigma_3) = (8, 6, 4)^T$, $(10, 8, 6)^T$, $(12, 10, 8)^T$, $(14, 12, 10)^T$ or $(18, 14, 12)^T$. Under these five sets of noise levels, the top panel compares the true trivariate β (red horizontal lines) with the boxplot of its estimate and the bottom panel displays the same result but for $(\sigma_1, \sigma_2, \sigma_3)$. Overall the noise variances are well estimated while the β's have a slight bias when the latter is large. In addition, the noise level does not greatly affect the quality of our estimation.

Figure 5.9 is the same as in Figure 5.8 but with a fixed $(\sigma_1, \sigma_2, \sigma_3) = (1.0, 1.0, 1.0)^T$ and five different $\pi = 0.005$, 0.010, 0.015 ,0.020, or 0.025. These graphs show that changing the number of change-points, i.e. driven by π, does not have a strong effect on the estimation of the β's and of the σ_i's.

5 WAM Results and discussion

Our statistical model and inference method are now applied to the bivariate vector composed of the OLR and ITD time series described in Section 3 for each year starting in 1979 and ending 2008. To interpret these outputs, we focus our attention on the four years (1990, 1992, 1998 and 2006) introduced in Figure 5.4. The top and middle panels of Figures 5.10, 5.11, 5.12 and 5.13 show, in red, the addition of the estimated trends $(f_j(t))$ and the extracted break signals $(\beta_j x_t)$ of Equation (5.2), for the OLR and the ITD, respectively. A visual inspection tends to indicate that the trends are well estimated. Note that, the trends $(f_j(t))$ do not have a physical meaning, here. Nevertheless their estimations are necessary to successfully detect onset dates, as the estimations of both signals $(f_j(t))$ and x_t of equation (5.2) have to be calculated simultaneously. Indeed, statistical methods generally require stationnary time series to be reliable. The calculation of $f_j(t)$ using an autoregressive spline estimation (e.g. equation (5.5) can be considered as the stationarisation process of our method. In other applications, as for instance homogenization problems (i.e. the detection of artificial shifts in time series, e.g. Caussinus and Mestre [2004]), these trends would have physical meaning. Concerning the change-points estimation, the extracted common signals seem to be reasonable. For example, Figure 5.11 clearly indicates an onset around the end of June 1992 and spurious change-points appear in the Spring and October 1992. Those latter changes are due to poor data quality (missing

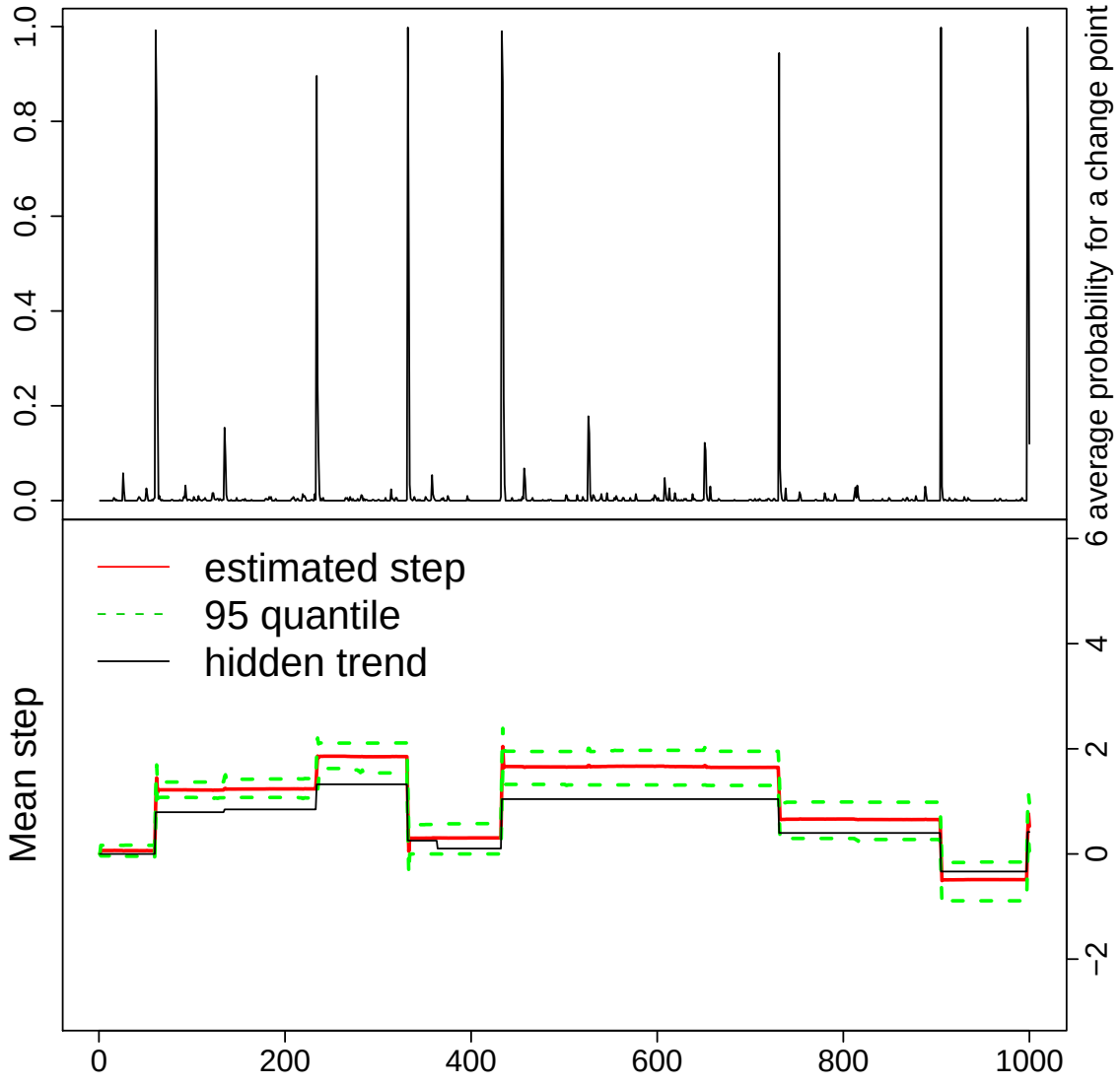

FIGURE 5.7 – The top panel represents the estimated probability of observing change-points simultaneously in the three time series displayed in Figure 5.6. The bottom panel compares the true (black) x_t defined by (5.3) and its estimate (red) with their 95% confidence interval (dotted green lines).

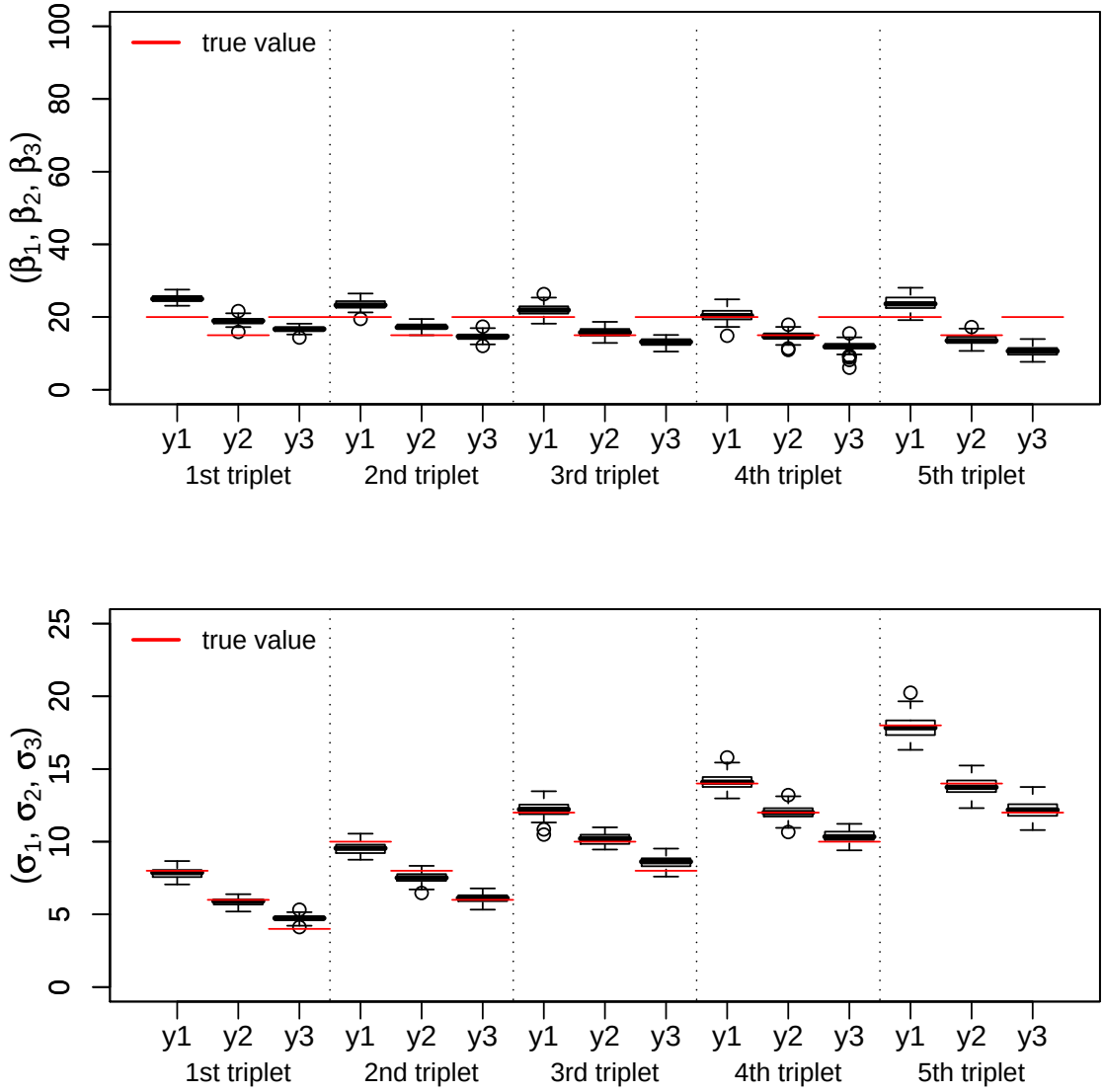

FIGURE 5.8 – Boxplots from 500 simulations with $\beta^T = (20, 15, 20)^T$ and $\pi = 0.01$ and the trends of Figure 5.6. The x-axis corresponds to five different combinations of the triplet $(\sigma_1, \sigma_2, \sigma_3)^T = (8, 6, 4)^T$, $(10, 8, 6)^T$, $(12, 10, 8)^T$, $(14, 12, 10)^T$ or $(18, 14, 12)^T$. Under these five sets of noise levels, the top panel compares the true trivariate β (red horizontal lines) with the boxplot of its estimate and the bottom panel displays the same result but for $(\sigma_1, \sigma_2, \sigma_3)$.

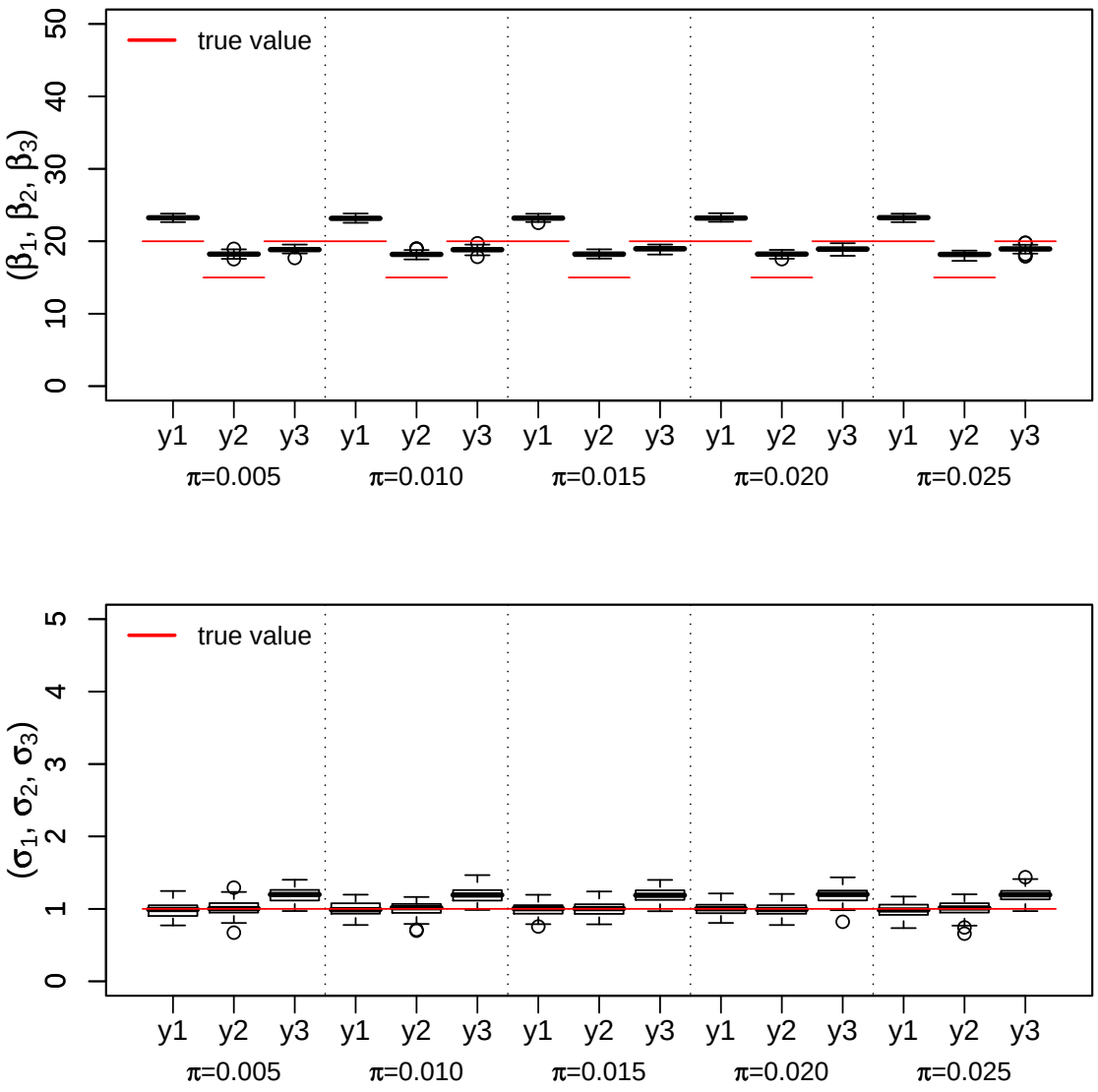

FIGURE 5.9 – Same as in Figure 5.8 but with a fixed $(\sigma_1, \sigma_2, \sigma_3)^T = (1.0, 1.0, 1.0)^T$ and five different $\pi = 0.005$, 0.010, 0.015 ,0.020, or 0.025.

data, edge effects) and should be disregarded as obvious artefacts in the context of WAM onsets. The same type of reasoning can be employed for the years 1990, 1998 and 2006. Both, a pre-onset (in June) and an onset (in July) can be easily identified for years like in Figure 5.10, while this is not possible for others, see 1992.

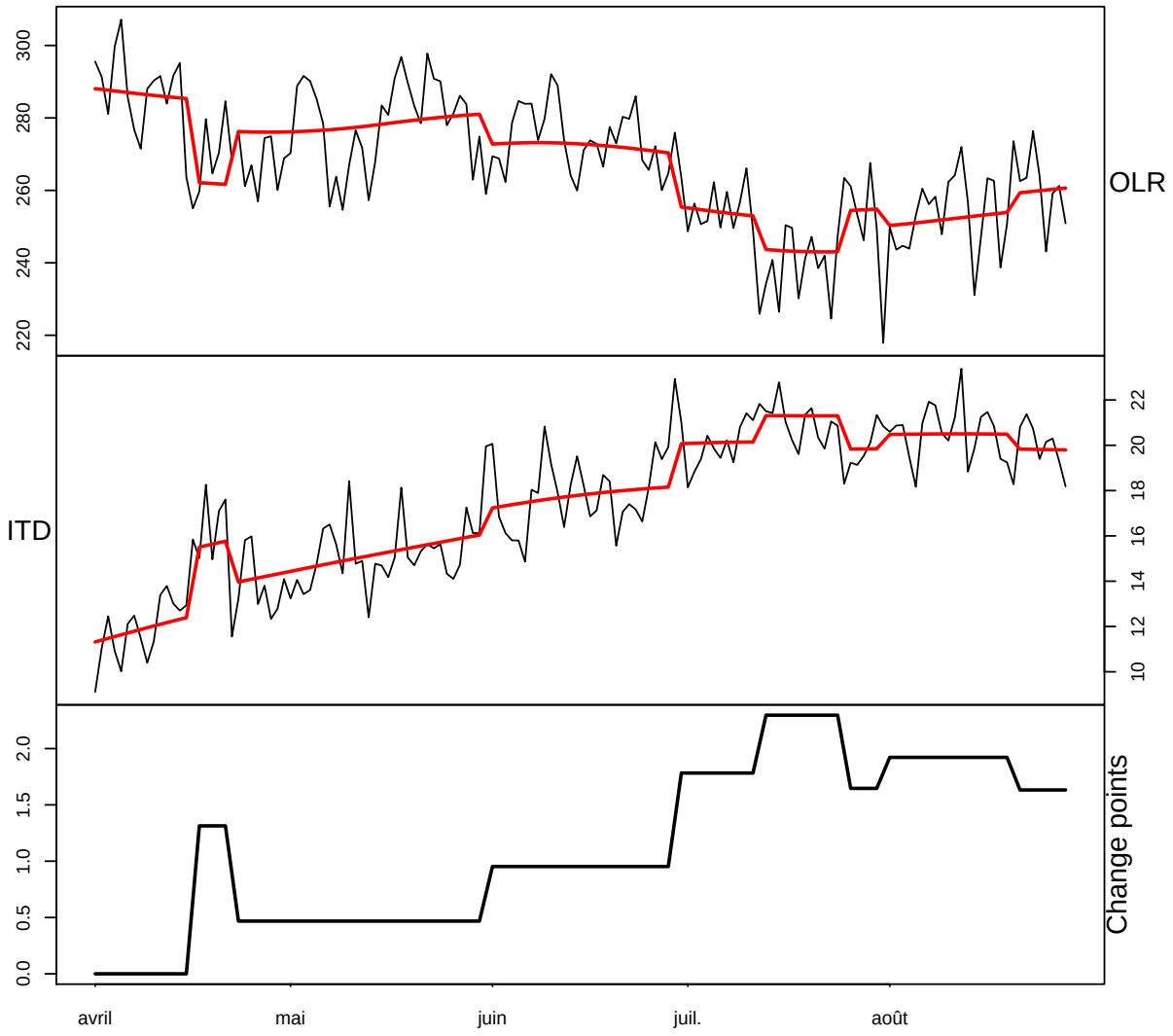

FIGURE 5.10 – Statistical treatment of the 1990 OLR and ITD times series from the top panel of Figure 5.4. The red line corresponds to the estimated trend $f_1(t)$ and $f_2(t)$ from Equation (5.2). The bottom panel displays the extracted hidden change-point signal x_t from Equation (5.3).

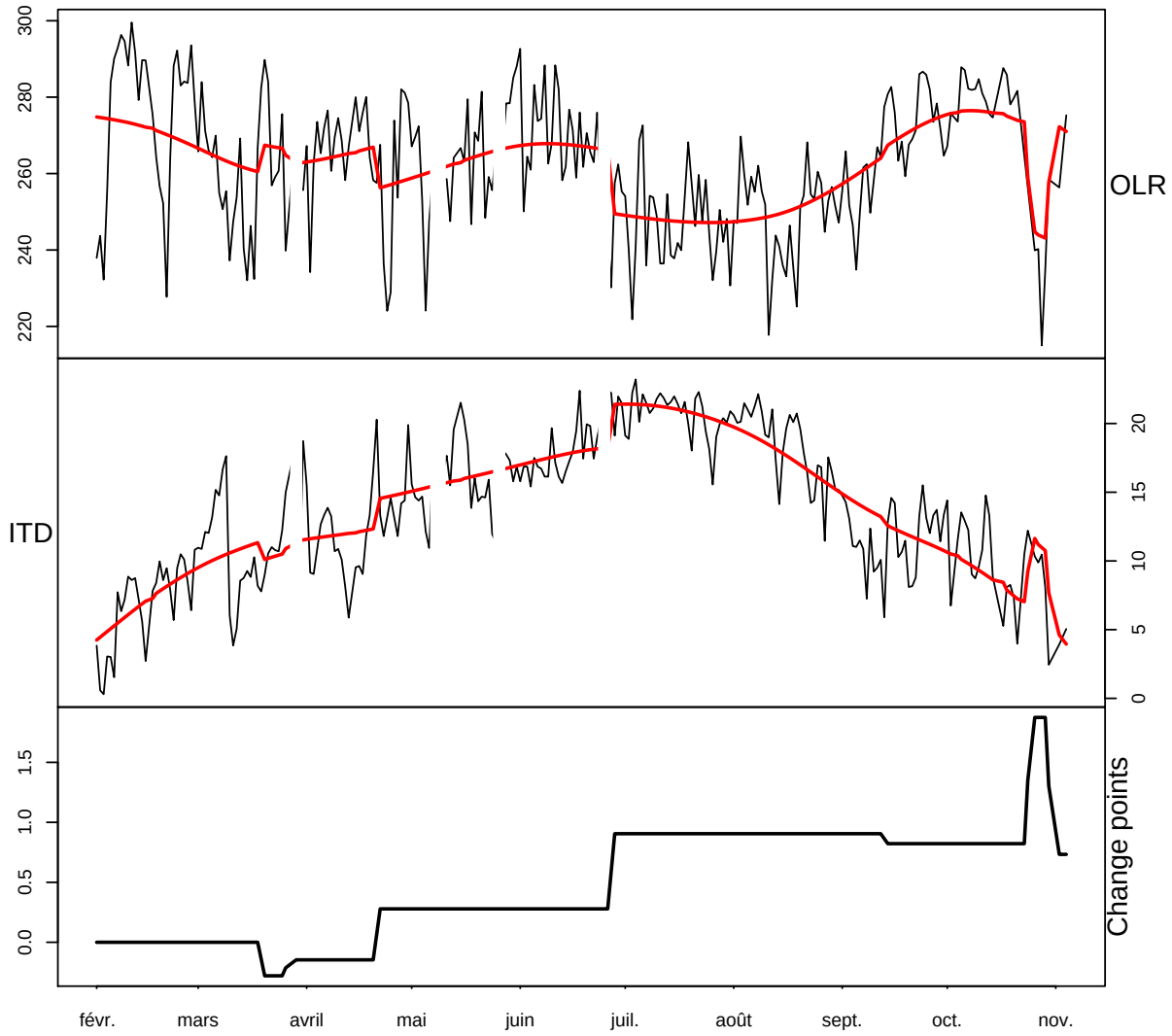

FIGURE 5.11 – Statistical treatment of the 1992 OLR and ITD times series from the second panel of Figure 5.4. The red line corresponds to the estimated trend $f_1(t)$ and $f_2(t)$ from Equation (5.2). The bottom panel displays the extracted hidden change-point signal x_t from Equation (5.3).

For some years like 1988 and 1991, we do not detect any significant onset because we do not force our model to find a specific number of change-point. We believe that is a strength, “the data speak from themselves", without a strong a priori on the yearly change-point number and consequently, if the time series are too noisy or the onset is too weak, then there is no detection.

The whole detected break points are illustrated in Figure 5.14. Each histogram bar represents the number of detected change points per day along the year with probability of occurrence $q_t^1 > 0.5$. The figure displays a bimodal density (black smooth line calculated as a kernel density, cf Parzen [1962]). The first mode corresponds to both onset and pre-onset signal mixing together around June, the second mode corresponds to the end of the monsoon season.

We are more particularly interested in the first mode and discriminate pre-onset from onset signal thanks knowledge from previous work on WAM (i.e. Sultan and Janicot [2003]). Some detected change points (as the one occurring in April 1992 see Figure 5.11) can be considered as spurious change points (i.e. not an onset signal), see Sultan and Janicot [2003].

Table 5.1 compares our results with the two different estimated dates by Fontaine et al. [2008]. Although not directly comparable because they were not derived from the same data, most of our dates fall between the proposed dates from Fontaine et al. [2008] or differ for about a week.

Figure 5.15 shows the frequency of our estimated WAM pre-onset and onset dates for the period 1979-2008. The former dates occur around the beginning of June and the latter around the beginning of July. The onset dates occurring on average on June 30th (with a standard deviation equal to 10 days) are consistent with the climatological date of June 24th found by Sultan and Janicot [2003]. Those authors have detected the central date of the transitional period, in contrast to our analysis which reveals the beginning of the post-onset period. The pre-onset dates as determined by the authors (May 14th) do not seem to be accurate with the average date of June 2nd (with a standard deviation equal to 8 days) found in this study. The reasons of this inconsistency should be highlighted. In Sultan and Janicot [2003] the pre-onset date is determined only by the ITD location. Our results on the pre-onset date show difficulties to capture this event, i.e. an abrupt northward propagation of the WAM before its onset. The pre-onset dates are mostly associated to a rather "anomalous" climatological cycle of deep convection over west Africa. For all years the pre-onset period is well defined presenting the ITCZ over the Guinean coast. However, for some years the WAM onset comes after none or even two transitional periods with intermediate phases being embedded. These intermediate phases are characterised

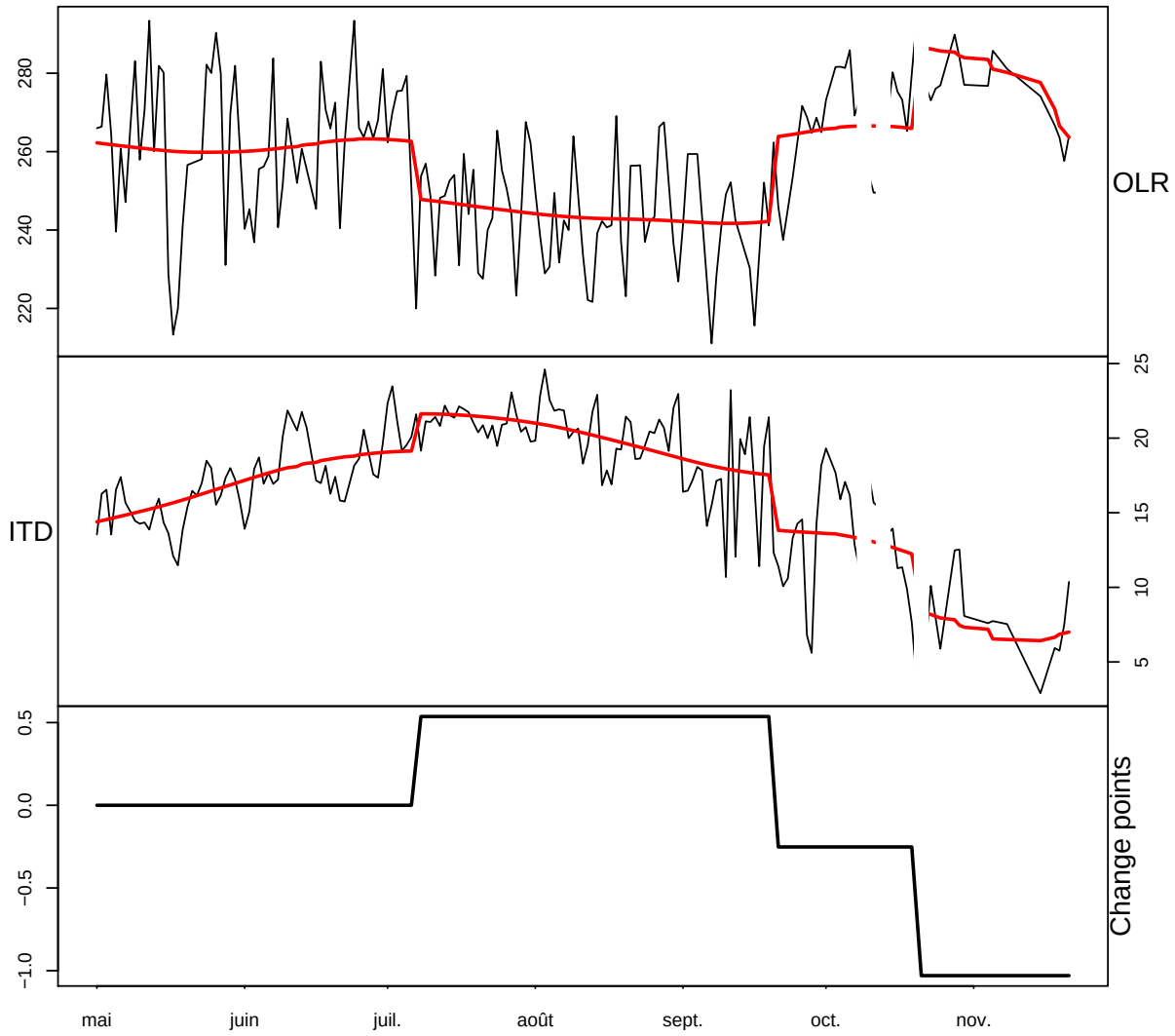

FIGURE 5.12 – Statistical treatment of the 1998 OLR and ITD times series from the third panel of Figure 5.4. The red line corresponds to the estimated trend $f_1(t)$ and $f_2(t)$ from Equation (5.2). The bottom panel displays the extracted hidden change-point signal x_t from Equation (5.3).

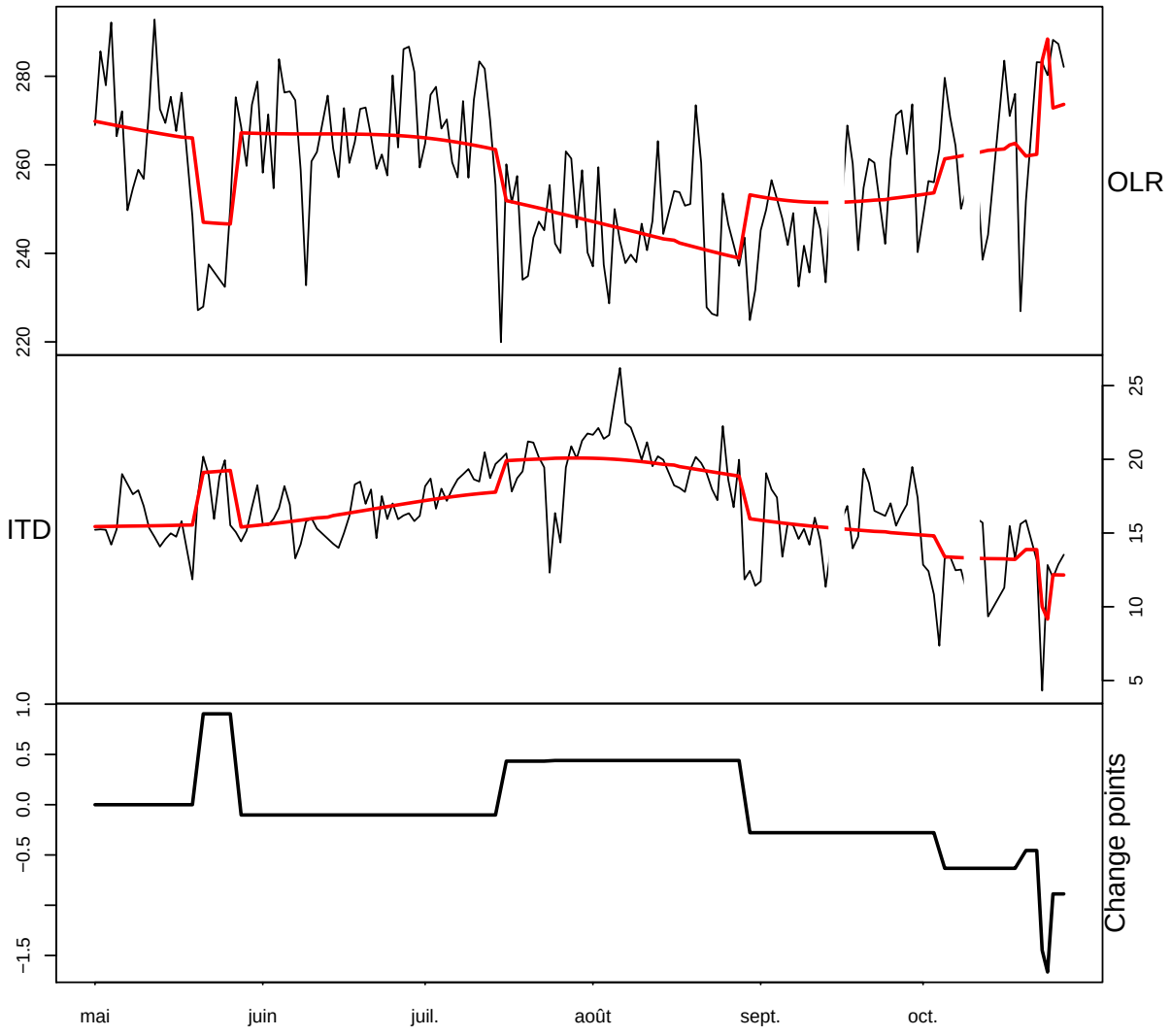

FIGURE 5.13 – Statistical treatment of the 2006 OLR and ITD times series from the bottom panel of Figure 5.4. The red line corresponds to the estimated trend $f_1(t)$ and $f_2(t)$ from Equation (5.2). The bottom panel displays the extracted hidden change-point signal x_t from Equation (5.3).

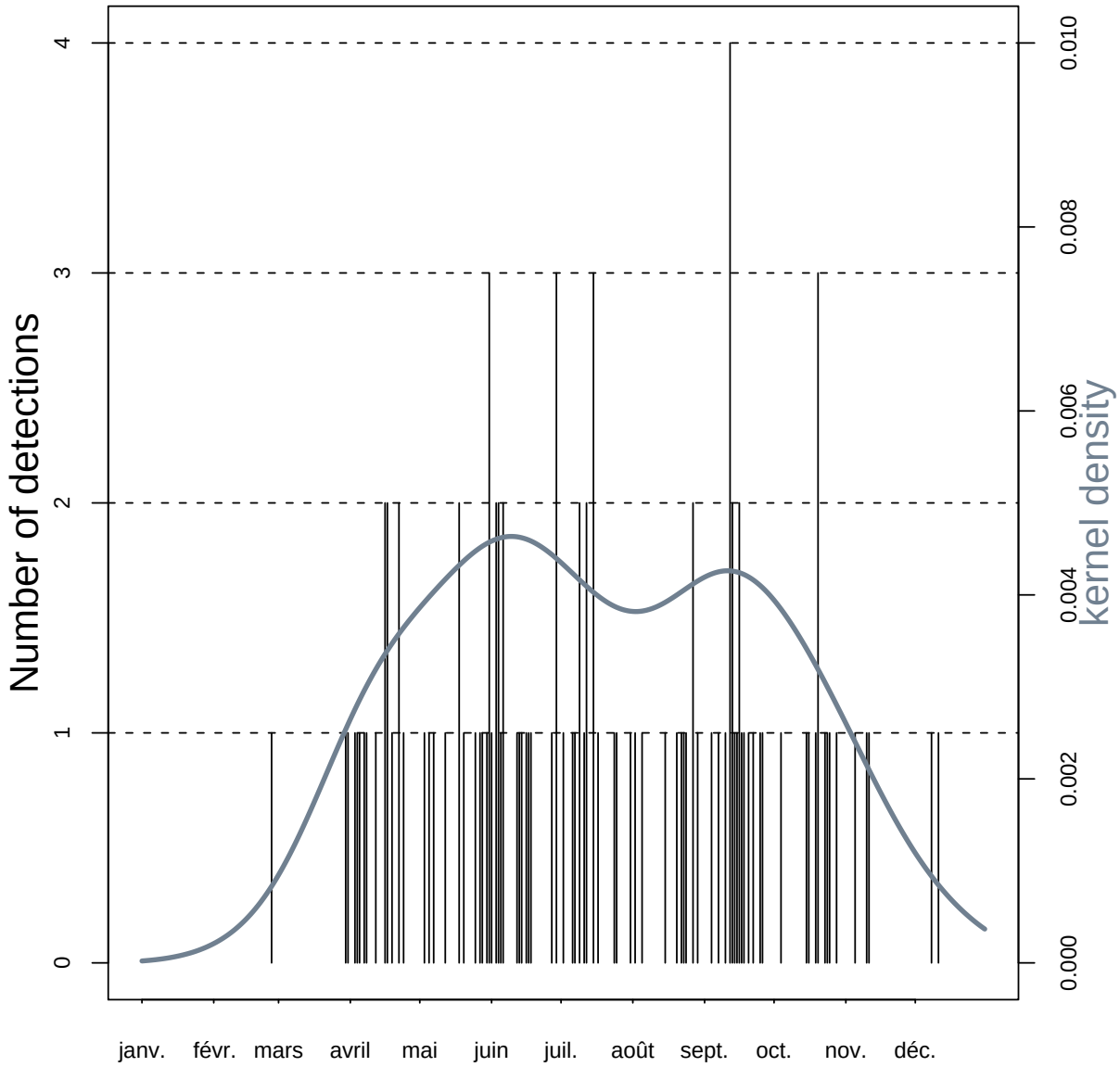

FIGURE 5.14 – The whole detected change points of each year of WAM times series from 1979 to 2008 with $q_t^1 > 0.5$. The smooth lines represent the density probability calculated with a Gaussian kernel as explained in Parzen [1962].

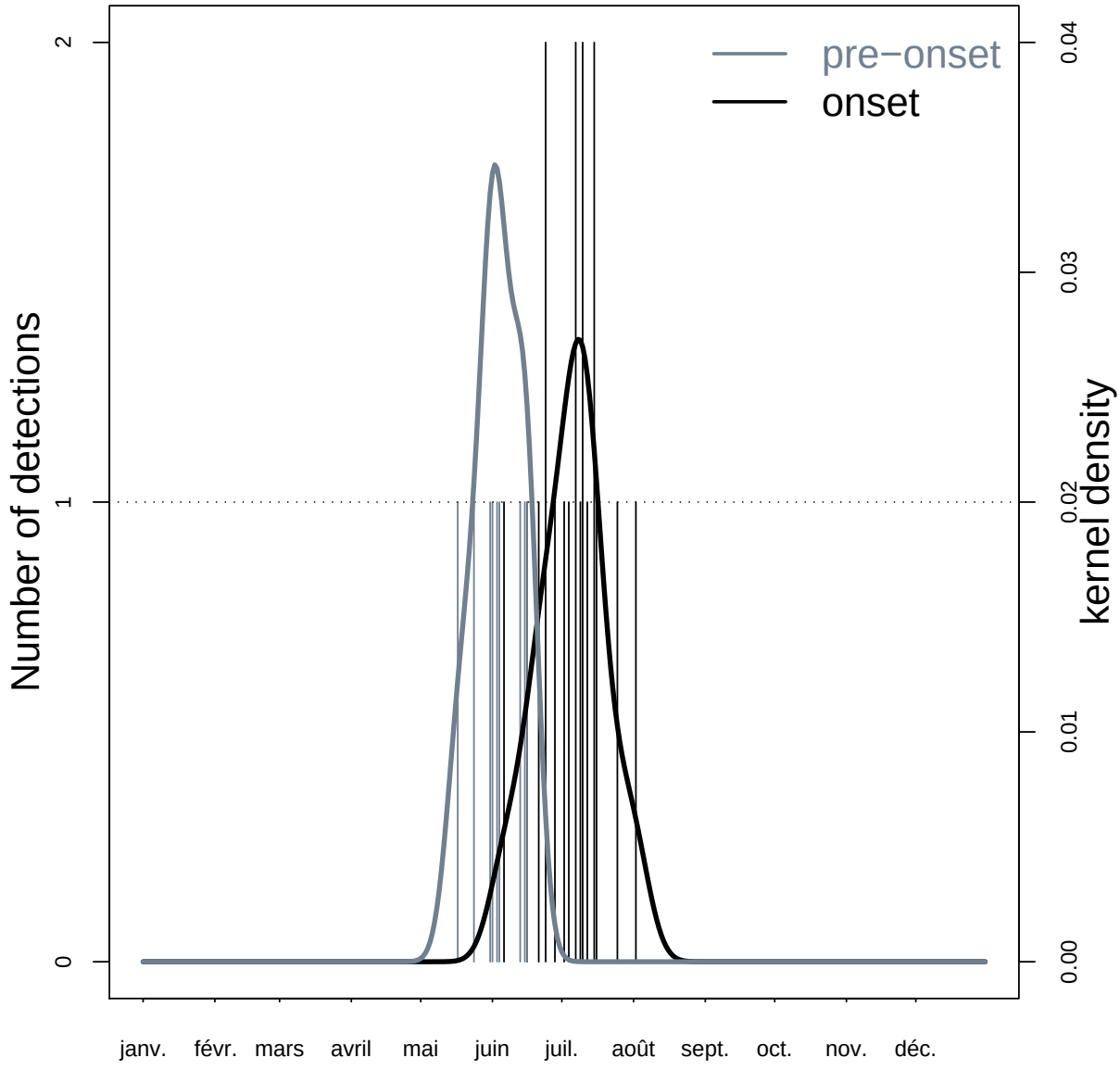

FIGURE 5.15 – The frequency of our estimated WAM pre-onset and onset dates for the period 1979-2008. The grey and black colours correspond to pre-onset dates occurring around the beginning of June and onset dates around the beginning of July, respectively. The smooth lines represent the density probability calculated with a Gaussian kernel as explained in Parzen [1962].

TABLE 5.1 – Comparison between our detected onset dates and the ones of Fontaine et al. [2008]

	1979	1980	1981	1982	1983	1984
Gazeaux et al.	07/10	06/26	06/25	06/05	-	07/10
Fontaine 08(1)	06/17	06/16	07/02	06/22	06/22	07/11
Fontaine 08(2)	07/22	07/01	07/12	07/01	07/02	08/05
	1985	1986	1987	1989	1990	1991
Gazeaux et al.	-	06/06	07/09	06/16	06/29	-
Fontaine 08(1)	06/22	06/22	06/27	07/01	06/26	07/11
Fontaine 08(2)	06/17	07/07	07/02	07/22	07/11	08/05
	1992	1993	1994	1995	1996	1997
Gazeaux et al.	07/06	-	-	06/21	06/14	07/25
Fontaine 08(1)	06/20	06/21	06/21	06/16	06/12	06/06
Fontaine 08(2)	06/25	06/26	07/01	06/16	06/12	06/06
	1998	1999	2000	2001	2002	2003
Gazeaux et al.	07/07	06/28	07/03	06/16	07/15	07/10
Fontaine 08(1)	07/01	07/11	06/25	07/01	06/16	06/16
Fontaine 08(2)	07/01	09/01	07/05	07/11	08/05	06/26

	2004	2005	2006	2007	2008
Gazeaux et al.	-	06/10	07/15	07/02	06/23
Fontaine 08(1)	07/05	-	-	-	-
Fontaine 08(2)	06/15	-	-	-	-

by the presence of convective activity over both the Guinean coast and the Sahel. This might create confusion on the WAM onset detection. If no abrupt signal is detected (due to a smoothed ITCZ cycle) by the model then the WAM onset may not be defined (e.g. in 1993) or if intermediate phases are present then the model may detect more monsoon jumps (e.g in 2002). In that case two dates are defined : a "pre-onset" and an "onset" date. These intermediate phases may be the reason for the large difference of the onset dates for some years, as in 2002, between this study and the study of Fontaine et al. [2008]

Finally, to conclude this paper, we would like to say a few words of caution. The detection of the onset appears then to be complex. Under the definition of the onset by Sultan and Janicot [2003], we expected to find a single clear change point occurring around the end of June, but our methodology clearly detects more than one change point per year. The differences open new questions. Is there really a unique date for the yearly onset ? Are our OLR and ITD data the most appropriate time series for detections ?

Convection over West Africa is a result of complex dynamics and different forcings from regional or larger scale climate. Hence, the evolution of the localisation of convection could be considered somewhat independent of the ITD. Our statistical approach on

the WAM onset could be optimized by using other elements of the west African climate associated to convection and thus eliminating any "false onset". This issue is a perspective for future studies.

To summarize this article, we recall the reader that our objective was to propose a non-linear statistical extraction method that can both infer individual smooth trends and common change-points in multivariate time series. From the simulation study, it appears that the proposed inference procedure based on Kalman filtering ideas works adequately. We illustrate the applicability of our method by detecting pre-onset and onset dates of convection over the Sahel related to the WAM. For this specific application, the advantage of our approach resides in the global representation of uncertainties and it does not contradict similar studies based on different data and simpler statistical techniques.The estimation of the onset dates distribution of Figure 5.15 could also likely be treated as relevant a priori information for future prediction studies. The generic aspect of our modeling strategy could be exploited for other climate studies that focus on differencing smooth trends and abrupt discontinuities. We discussed the adequation of our method for homogenization problem, of course, our assumption that change points have to occur simultaneously in time could be a limitation in homogenization. Future research is needed to modify our algorithm in order to tailor it to a specific homogenization case study.

6 Appendices

6.1 The calculations of the Non linear Kalman Smoother

For information on the calculations below, we inspired on different books such as books by Hossack et al. [1999] or Basseville and Nikiforov [1993].
To simplify the writing, we are using the obvious following notations :
$\hat{X}(t|Y_{1:t}) \doteq \mathbb{E}[X_t|Y_{1:t}]$ for the expectation of X_t conditioned on $Y_{1:t}$, and $\hat{\Sigma}(t|Y_{1:t}) \doteq \mathbb{V}ar[X_t|Y_{1:t}]$ for the variance of X_t conditioned on $Y_{1:t}$.
1st step : prediction step - The first step of the KF solution begins with the estimation of the prediction. It means, the calculation of the recurrence relation between $\hat{X}(t|Y_{1:t-1}, b_t)$ and $\hat{X}(t-1|Y_{1:t-1}, b_t)$, and also the similar relation for the variance $\hat{\Sigma}(t|Y_{1:t-1}, b_t)$ and $\hat{\Sigma}(t-1|Y_{1:t-1}, b_t)$
$\forall i = 0, 1$

$$\begin{aligned}\hat{X}(t|Y_{1:t-1}, b_t = i) &= \mathbb{E}[\Phi X_{t-1} + E_t|Y_{1:t-1}, b_t = i] \\ &= \Phi\mathbb{E}[X_{t-1}|Y_{1:t-1}, b_t = i] \\ &\quad + \mathbb{E}[E_t|Y_{1:t-1}, b_t = i] \\ &= \Phi\hat{X}(t-1|Y_{1:t-1}) + W_i \end{aligned} \tag{5.12}$$

and

$$\hat{\Sigma}(t|Y_{1:t-1}, b_t = i) = \Phi\hat{\Sigma}(t-1|Y_{1:t-1})\Phi' + \mathbb{C}ov(W_i)$$

where
$W_0 = [0 \,..\, 0]$, $W_1 = [\mu\ \mu\ 0 \,..\, 0]$,

$$\mathbb{C}ov(W_i) = \begin{bmatrix} \Gamma_i & 0 & 0 & 0 \\ 0 & \mathbb{C}ov(E_{f_j}) & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & Cov(E_{f_j}) \end{bmatrix},$$

$$\mathbb{C}ov(E_{f_j}) = \lambda_j\sigma_j^2 \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{bmatrix}$$

$$\Gamma_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \text{ and } \Gamma_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

2nd step - Here is the calculation of the predicted distribution of the observations at time t conditioned on the available observations at this time and the occurrence of a change-point.

$\forall i = 0, 1$

$$\begin{aligned} \mathbb{E}[Y_t|Y_{1:t}, b_t = i] &= \mathbb{E}[HX_t + E_t|Y_{1:t-1}, b_t = i] \\ &= H\hat{X}(t|Y_{1:t}, b_t = i) \\ \mathbb{V}ar[Y_t|Y_{1:t-1}, b_t = i] &= H\hat{\Sigma}(t|Y_{1:t-1}, b_t = i)H' \\ &\quad + \mathbb{V}ar(E_t) \end{aligned} \tag{5.13}$$

We can estimate distribution at time t with the same distribution but also conditioned on the occurrence of a break by the approximate mixture of multivariate normal distribution :

$$\begin{gathered} (Y_t|Y_{1:t}) \text{ is equal in distribution to} \\ (1-\pi)(Y_t|Y_{1:t}, b_t = 0) + \pi(Y_t|Y_{1:t}, b_t = 1) \end{gathered} \tag{5.14}$$

3rd step : Calculation of the probability q_t^i - This calculation provides the posterior probability at every time of the occurrence of a change-point conditioned on the observations available at this same time $Y_{1:t}$.

$$\begin{aligned} q_t^0 \doteq Pr(b_t = 0|Y_{1:t}) &= \frac{(1-\pi)Pr(Y_t|Y_{1:t-1}, b_t = 0)}{Pr(Y_t|Y_{1:t-1})} \\ q_t^1 \doteq Pr(b_t = 1|Y_{1:t}) &= \frac{\pi Pr(Y_t|Y_{1:t-1}, b_t = 1)}{Pr(Y_t|Y_{1:t-1})} \end{aligned} \tag{5.15}$$

4th step : update state - This step is the second main of the KF. It deals with the update of the dynamics with the observations of the current time. We have to express a relation between $\hat{X}(t|Y_{1:t}, b_t = i)$ and $\hat{X}(t|Y_{1:t-1}, b_t = i)$ and a similar relation for the second order : $\hat{\Sigma}(t|Y_{1:t}, b_t = i)$ and $\hat{\Sigma}(t|Y_{1:t-1}, b_t = i)$.
$\forall i = 0, 1$

$$\begin{aligned} \hat{X}(t|Y_{1:t}, b_t = i) &= \hat{X}(t|Y_{1:t-1}, b_t = i) \\ +\hat{\Sigma}(t|Y_{1:t-1}, b_t = i)H'\mathbb{V}ar[Y_t|Y_{1:t-1}, b_t = i]^{-1}[Y_t - \mathbb{E}[Y_t|Y_{1:t-1}, b_t = i]] \\ \hat{\Sigma}(t|Y_{1:t}, b_t = i) &= \hat{\Sigma}(t|Y_{1:t-1}, b_t = i) \\ -\hat{\Sigma}(t|Y_{1:t-1}, b_t = i)H'\mathbb{V}ar[Y_t|Y_{1:t-1}, b_t = i]^{-1}H\hat{\Sigma}(t|Y_{1:t-1}, b_t = i) \end{aligned} \tag{5.16}$$

5th step - Finally we introduce the probability q_t^i to take into account the non linearities of the occurrence of a change-point at time t. This second order of this step is achieved by using some well-known results on conditional variance, described in Hossack et al. [1999].

$$\begin{aligned}\hat{X}(t|Y_{1:t}) &= q_t^0 \hat{X}(t|Y_{1:t}, b_t = 0) + q_t^1 \hat{X}(t|Y_{1:t}, b_t = 1) \\ \hat{\Sigma}(t|Y_{1:t}) &= \sum_{i=0}^{1} \hat{\Sigma}(\mathbb{E}[X|Y_{1:t}, b_t = i]) + \mathbb{E}(\hat{\Sigma}[X|Y_{1:t}, b_t = i]) \\ &= \sum_{i=0}^{1} q_t^i \hat{\Sigma}(t|Y_{1:t}, b_t = i) \\ &\quad + q_t^i [\hat{X}(t|Y_{1:t}, b_t = i) - \hat{X}(t|Y_{1:t})]^2\end{aligned} \tag{5.17}$$

6th step : the Fixed Interval Smoother - Also called the Kalman filtering, this method permits to reconstruct the different components of the state vector given the entire time series, i.e. OLR or ITD). for all t from 1 to T, the FIS constructs $(X_t|Y_{1:T})$. We can obtain the Equation (5.10-5.19)

$$\begin{aligned}\hat{X}(t|Y_{1:T}) &= \hat{X}(t|Y_{1:t}) + C_t[\hat{X}(t+1|Y_{1:T}) - \hat{X}(t+1|Y_{1:t})] \\ \hat{\Sigma}(t|Y_{1:T}) &= \hat{\Sigma}(t|Y_{1:t}) + C_t[\hat{\Sigma}(t+1|Y_{1:T}) - \hat{\Sigma}(t+1|Y_{1:t})]C_t'\end{aligned} \tag{5.18}$$

where

$$C_t = \hat{\Sigma}(t|Y_{1:t})\Phi\hat{\Sigma}(t+1|Y_{1:t})^{-1} \tag{5.19}$$

The algorithm underlying this method was made with the free software environment for statistical computing and graphics "R". "R" provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. For more details see R Development Core Team [2009].

The authors acknowledge the support of the LATMOS((Laboratoire Atmosphères, Milieux, Observations Spatiales, www.latmos.ipsl.fr) a geosciences laboratory belonging to the CNRS/IPSL, and also the support of the GEOmon project : www.geomon.eu. The programs used were made with the functional language and environment R : www.r-project.org, see R Development Core Team [2009]. Part of this work has been suppor-

ted by the EU-FP7 ACQWA Project (www.acqwa.ch) under Contract Nr 212250, by the PEPER-GIS project and by the ANR-MOPERA project.

7 Validation a-posteriori de la méthode : caractéristiques des résidus

À l'instar de l'étude des résidus faite au chapitre 4, nous présentons, dans cette section quelques résultats sur les résidus, $\epsilon_j(t)$, obtenus grâce à notre méthode de décomposition du modèle (5.1). Ces résultats montrent le bien fondé du développement de ce chapitre. Nous opérons tout d'abord sur les données simulées présentées dans l'article, puis sur les données de l'application. Nous montrons ainsi que les hypothèses faites afin de résoudre les équations du filtre de Kalman restent vérifiées à la sortie du modèle.

7.1 Résidus des données simulées

Les Figures 5.16, 5.17 et 5.18 présentent l'étude des résidus obtenus à partir des données simulées. Les résidus obtenus présentent des distributions Gaussiennes (cf *QQ-plots* de la Figure 5.17), ainsi qu'une indépendance temporelle illustrée par les fonctions d'autocorrélation (Figure 5.18). Ces deux hypothèses émises pour résoudre les équations du filtre de Kalman sont ainsi bien vérifiées sur les résidus extraits à partir de données simulées.

7.2 Résidus des données de l'application

Les Figures 5.19, 5.20 et 5.21 présentent les résultats sur les résidus obtenus à partir des données de l'application à la détection de la Mousson de l'Afrique de l'ouest. La première série représente les données d'OLR, la seconde les données d'ITD. On peut remarquer que les séries suivent une distribution assimilables à une distribution gaussienne (Figure 5.20) et que les la fonction d'autocorrélation semble présenter une faible dépendance d'ordre 1, qui, cependant, ne semble pas significative (Figure 5.21).

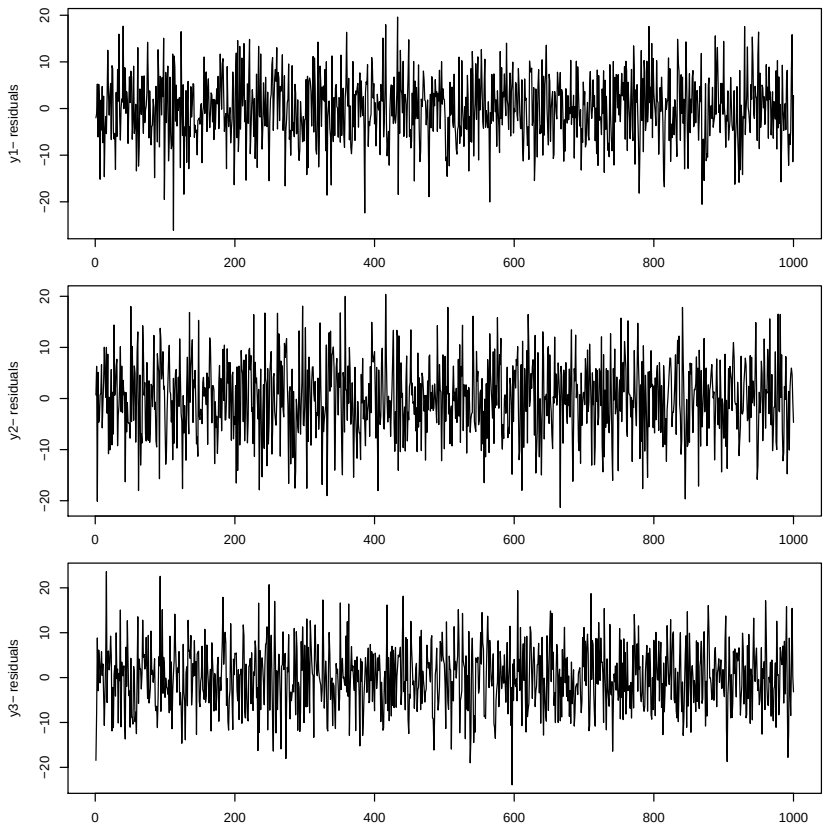

FIGURE 5.16 – Séries brutes des signaux $\epsilon_j(t)$ issus de la décomposition des séries présentées à la Figure 5.6.

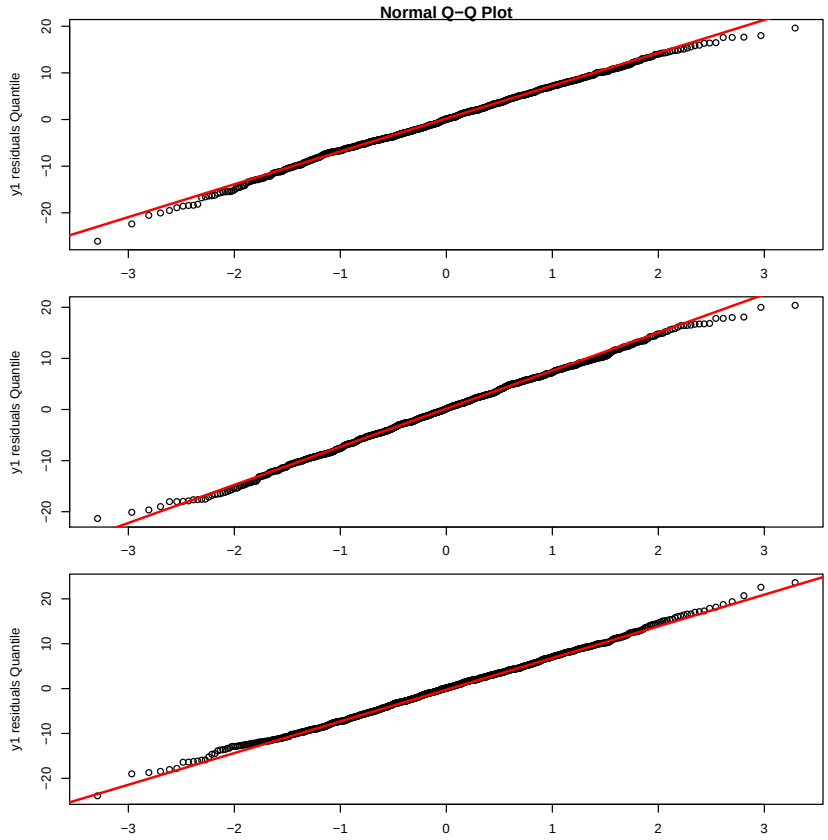

FIGURE 5.17 – *QQ-plots* des séries $\epsilon_j(t)$ de la Figure 5.16. On remarque que la distribution des résidus (en noir) est assimilable à une distribution Gaussienne (en rouge).

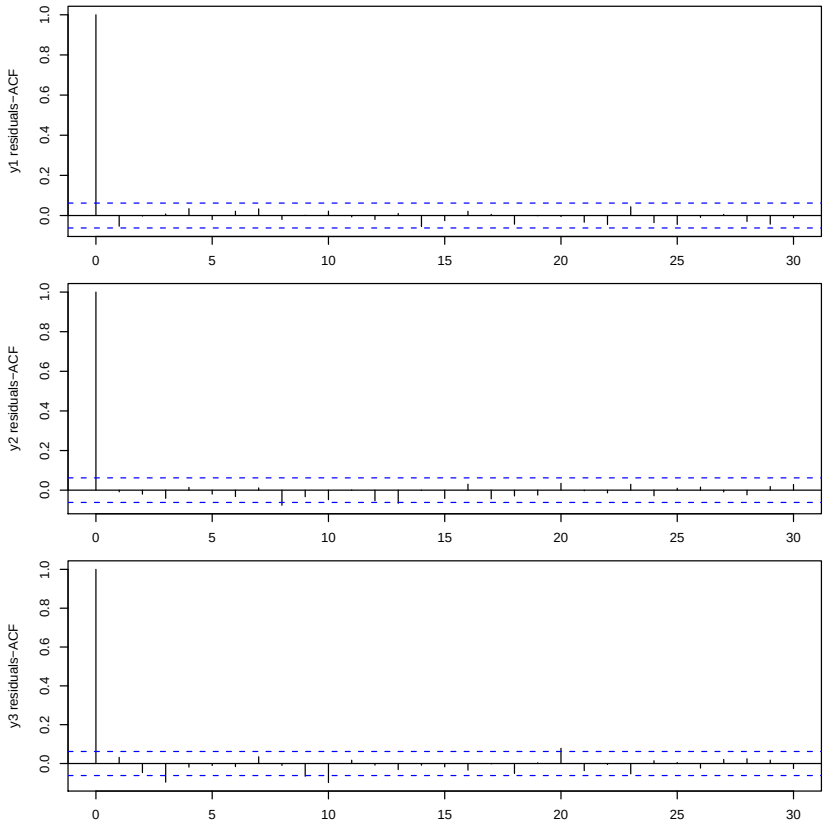

FIGURE 5.18 – Fonction d'autocorrelation des séries $\epsilon_j(t)$ de la Figure 5.16. On peut considérer que, en dehors de l'ordre 0, les autocorrélations du signal sont nulles. Ceci illustre, pour chaque série j considérée, le caractère indépendant des différents $\epsilon_j(t)$.

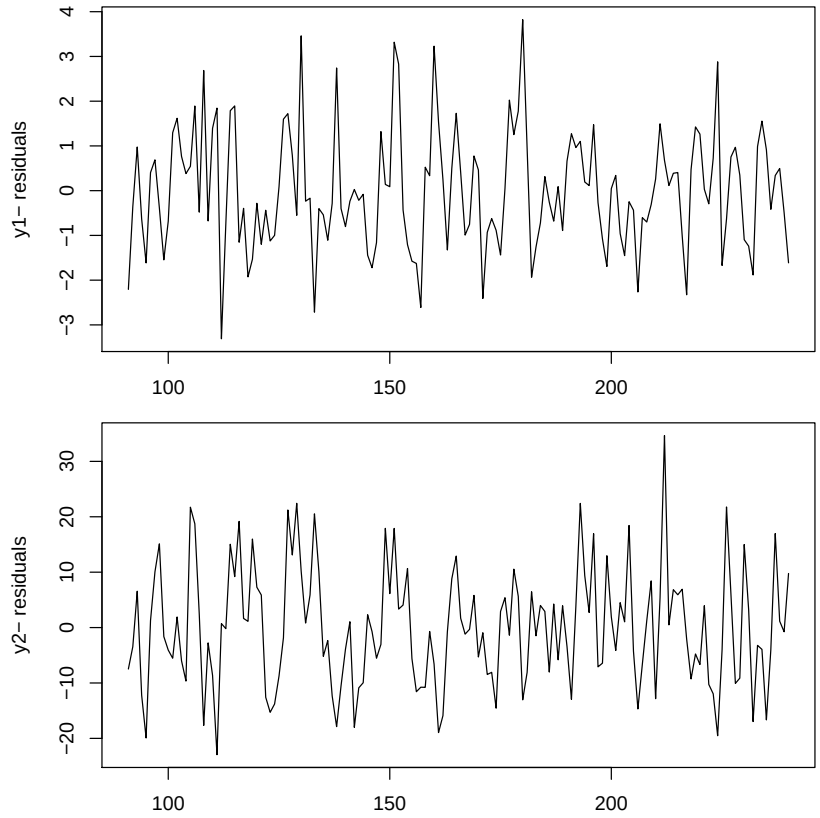

FIGURE 5.19 – Signaux des résidus issus de l'extraction sur les données de l'ITD et de l'OLR.

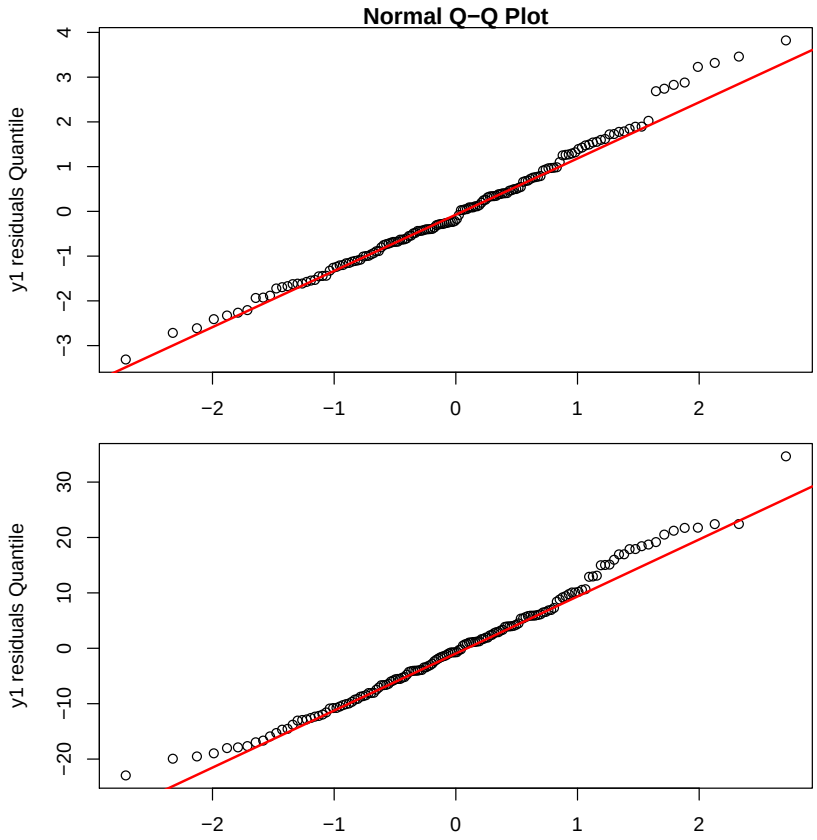

FIGURE 5.20 – Présentations des *QQ-plots* des séries de la Figure 5.19. Les résidus (en noir) présentent clairement une distribution Gaussienne (en rouge).

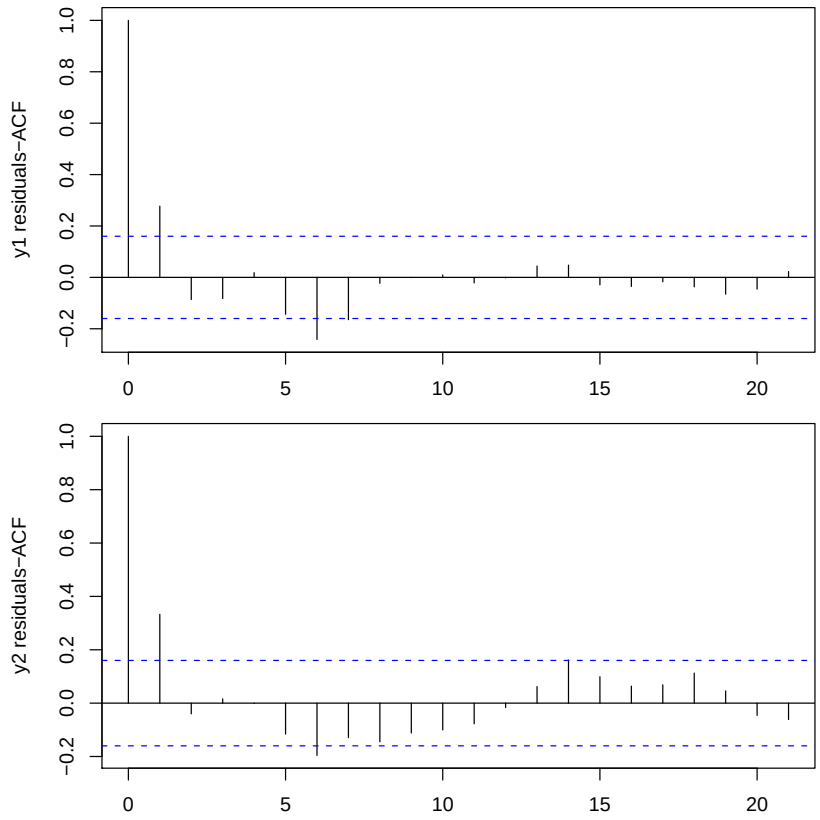

FIGURE 5.21 – Présentation de la fonction d'autocorrélation des séries de la Figure 4.12. Ici, les résidus présentent une faible dépendance d'ordre 1, qui ne semble pas significative.

Chapitre 6

Conclusion Générale

1 Retour sur les chapitres

Après avoir circonscrit les enjeux de l'extraction de signaux cachés et le cadre probabiliste qui s'y rattache, nous avons exploré, dans ce travail de thèse différents aspects de l'extraction de signaux dans des séries de données en sciences de l'atmosphère. Au travers de trois problématiques d'intérêt scientifique fort, nous avons construit des modèles d'inférence probabiliste adaptés à la décomposition de signaux et à la multitude d'observations disponibles. Ces modèles ont tout d'abord été évalués sur des données simulées avant d'être appliqués à des données réelles d'observation. Nous étudions enfin les résidus des modèles afin de tester les hypothèses émises pour résoudre ces problèmes. L'ensemble de ces développements méthodologiques a été réalisé grâce au logiciel de statistique R (voir R Development Core Team [2009]). Un package notamment a été construit (voir annexe C) avec documentation afin de diffuser ces outils.

Le premier problème considéré a été la détection de nuages stratosphériques polaires dans des profiles verticaux lidar de retrodiffusion mesurés à la station Antarctique de Dumont D'Urville. Il nous a amené à nous intéresser au traitement de données non stationnaires, hétéroscédastiques et au choix d'un modèle probabiliste décrivant le signal et ses différentes composantes. La couche de nuage est caractérisée par une rupture transitoire de variance du signal. Un calcul de maximisation d'un rapport de vraisemblance permet de statuer de manière objective sur l'existence d'une couche nuageuse dans des profils lidar. Un résultat important obtenu avec notre méthode de détection est la mise en évidence des effets induits par le moyennage des profils de rétrodiffusion, qui est une pratique courante lors de l'analyse de données lidar.

Le deuxième développement présenté s'est focalisé sur l'étude de signaux volcaniques cachés dans des séries temporelles de dépôt de sulfate extrait de carottes de glace provenant de différents forages au Groenland. Cette étude nous a amené à développer une méthode de décomposition de séries aléatoires grâce à un filtre de Kalman non stationnaire. Le modèle prend en compte la présence simultanée de signaux volcaniques non-linéaires caractérisés par un pic soudain suivi d'une décroissance relative à la diffusion et la disparition des sulfates dans l'atmosphère. À l'aide de notre algorithme de détection des signaux volcaniques, nous avons pu détecter avec une grande certitude les principales éruptions volcaniques depuis 1645 ainsi que des éruptions plus modestes, ayant des probabilités d'occurrence plus faibles.

Enfin, la dernière étude a porté sur le Mousson de l'Afrique de l'ouest, et plus précisément sur son instant de déclenchement. Ce déclenchement est caractérisé aux alentours de Juin/Juillet par une rupture soudaine de la dynamique atmosphérique régionale, à sa-

voir, le déplacement vers le Nord du front intertropical et l'amplification des pluies. Nous avons développé un modèle de détection de ruptures dans des séries temporelles des données ayant des comportements différents mais un signal caché commun. Le formalisme est proche de celui de l'étude précédente avec des séries multivariées. Néanmoins des problèmes sons apparus tels que les conséquences sur l'incertitude de l'estimation de la non stationnarité du signal caché que l'ont souhaite extraire. Cette étude a de nouveau été réalisée grâce à un filtre de Kalman, permettant, cette fois-ci la détection de ruptures ayant des instants et des amplitudes aléatoires inconnues. Les résultats obtenus confirment l'adéquation du modèle avec les données, et ont permis d'évaluer une distribution des instants d'apparition de l'onset de la mousson sur une période de trente ans (entre 1978 et 2008).

2 Perspectives

Les préambules des Chapitres 3, 4 et 5 décrivent de manière synthétique différentes perspectives envisageables pour chacun des développements présentés. De manière générale, les différents modèles probabilistes développés peuvent être soit assouplis soit être sujet à un certain nombre de développements méthodologiques afin d'améliorer les performances des algorithmes et d'étendre les domaines d'application.

Par exemple, il est envisageable suite aux résultats concluant l'étude des nuages stratosphériques polaires, que le choix de la maximisation du rapport de vraisemblance soit complété par une approche bayésienne, en utilisant les résultats de l'étude comme connaissance a priori du sujet. Effectivement, en s'inspirant les travaux de détection de rupture de variance de [McCulloch and Tsay, 1993] ou [Diard et al., 2003], et en adaptant, par exemple, à ce type de travaux, les approches bayésiennes de détection de [Hannart and Naveau, 2009], cela permettrait d'utiliser la connaissance a priori à la fois sur les hauteurs, les épaisseurs et la saisonnalité de nuages, ou encore sur le fait que les nuages se propagent généralement sur plusieurs profils lidar consécutifs et ainsi d'apporter une analyse probabiliste de la fonction de densité de probabilité des nuages.

En ce qui concerne la modélisation probabiliste, le filtre de Kalman s'est avéré être une approche très satisfaisante pour la résolution des problèmes posés aux Chapitres 4 et 5. La décomposition simultanée des différentes composantes de l'équation d'état est un atout considérable par rapport à la méthode utilisée au Chapitre 3, qui opère la décomposition par étape. Cette dernière solution supposant par exemple, que l'extraction du signal caché x de l'équation (3.1) dépend de l'extraction du signal m précédente. Il sera ainsi intéressant de tenter de résoudre ce dernier problème, en opérant globalement sur

la minimisation de l'erreur quadratique de l'estimation. Ainsi, comme nous l'avons fait pour les filtres de Kalman présentés dans cette thèse, il semble possible de combiner au modèle (3.1) un calcul de probabilité d'occurrence des signaux cachés basé sur la prise en compte d'une dépendance temporelle des états de x. En effet les états du modèle (3.1), à l'inverse des modèles (4.1) et (5.1) ne présente pas de mémoire auto régressive sur le signal caché. Ces développements reviennent à considérer des modèles ayant des dimensions plus élevées, on trouve dans la littérature différents articles pouvant nous guider sur cette voie ([Woods, 1981], [Hamill and Snyder, 2000]). On pourrait ainsi considérer un modèle plus développé du type :

$$P(z,t) = m(z) + x(z,t) + \epsilon(z),$$

et modéliser une dépendance temporelle entre les signaux $x(z,t)$ et $x(z,t-1)$, c'est à dire, évaluer comment se propage le signal lidar du nuage dans le temps.

Dans le même ordre d'idée, les développements des Chapitres 4 et 5 ne considèrent pas la dimension spatiale. On pourrait essayer de prendre en compte et modéliser la dépendance spatiale entre les états respectifs des modèles (4.1) et (5.1), ainsi qu'entre les bruits d'observation. Le modèle pourrait se décliner de la manière suivante :

$$y_j(t,r) = f_j(t) + \beta_j V(t,r)x(t) + \epsilon_j(t,r),$$

où r représente le vecteur de dimension spatiale, x le signal caché, et $V(t,r)$ modélise la dépendance spatiale du signal caché, et par l'intermédiaire de la matrice de covariance non diagonale $\epsilon_j(t,r)$ supposerait également une dépendance spatiale des bruits. On trouvera la présentation de différents modèles dans [Saporta and CoAuthors, 2008], et notamment certains chapitres de A. Montfort sur les filtres de Kalman et les variables cachées. Il sera également intéressant d'étudier plus en détail les travaux de thèse de Aurélien Ribes (e.g. [Ribes, 2009]) traitant de la détection de changements climatiques, grâce à l'analyse de modèles statistiques spatio-temporels et de différents développements proches de certains cités dans ce manuscrit (tests d'hypothèses, vraisemblance pénalisée ...).

Enfin, en ce qui concerne la condition contraignante de simultanéité des évènements dans les différentes séries d'observation (voir Chapitre 4 et 5), celle-ci pourrait être, à l'avenir relaxée et admettre des décalages temporels des évènements dus à la propagation spatiale des évènements et aux distances entre les différents sites d'observation ainsi qu'à des artefacts de mesure liés à l'incertitude sur certains types de données. Finalement, l'algorithme de détection de ruptures dans des séries, présenté au Chapitre 5 pourrait être

adapté à la détection de ruptures dans des séries non homogénéisées.

En ce qui concerne les projets nouveaux d'applications, nous avons envisagé des applications à la détection de couches volcaniques, de nuages mésosphériques polaires ou de turbulence en ciel clair qui se manifestent par une augmentation de la variance du signal de rétrodiffusion lidar en utilisant directement un modèle similaire à celui du Chapitre 3. Grâce au modèle présenté au Chapitre 4, nous envisageons aussi la détection d'oxyde d'azote produit par les éclairs (et ainsi, indirectement l'évaluation de la fréquence des éclairs dans une région donnée).

Bibliographie

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions : with formulas, graphs, and mathematical tables*. 1970.

A. Adriani, F. Cairo, L. Pulvirenti, F. Cardillo, M. Viterbini, G. Di Donfrancesco, and J. P. Pommereau. Stratospheric background aerosol and polar cloud observations by laser backscattersonde within the framework of the european project stratospheric regular sounding. *Annales geophysicae*, 17 :1352–1360, 1999.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 1974. doi : 10.1109/TAC.1974.1100705.

C. Ammann and P. Naveau. Statistical analysis of tropical explosive volcanism occurrences over the last 6 centuries. *GRL*, 30, 2003. doi : 10.1029/2002GL016388.

C. Ammann, G. A. Meehl, W. Washington, and C. Zender. A monthly and latitudinally varying volcanic forcing dataset in simulations of 20th century climate. *GRL*, 30, 2003. doi : 10.1029/2003GL016875.

C. Ammann, F. Joos, D. Schimel, B. Otto-Bliesner, and R. Tomas. Solar influence on climate during the past millennium : Results from transient simulations with the ncar climate system model. *Proc. Nat. Acad. Science*, 104 :3713–3718, 2007.

L. Amodei and J.-P. Dedieu. *Analyse numérique matricielle*. Dunod, 2008.

M. Basseville and I. V. Nikiforov. *Detection of abrupt changes : Theory and Application*. Prentice-Hall, Englewood Cliffs, N.J.,USA, 1993. ISBN 0-13-126780-9.

M. Basseville and I. V. Nikiforov. *Detection of abrupt changes : Theory and Application*. Prentice Hall, Englewood Clis. NJ, 1996.

C. Beaulieu, T. B. M. J. Ouarda, and O. Seidou. Synthese des techniques d'homogeneisation des series climatiques. *Hydrological Sciences Journal*, 52 :18–37, 2007. doi : 10.1623/hysj.52.1.18.

W. Bell. Signal extraction for nonstationary time series. *The Annals of Statistics*, 12 : 646–664, 1984.

S. Berthier, P. Chazette, J. Pelon, and B. Baum. Comparison of cloud statistics from spaceborne lidar systems. *Atmospheric Chemistry and Physics Discussions*, 8(2) :5269–5304, 2008. doi : 10.5194/acpd-8-5269-2008.

C. M. Bishop. *Information Theory, Inference, and Learning Algorithms*. Springer, 2006. ISBN 0387310738.

K. F. Boersma, H. J. Eskes, E. W. Meijer, and H. M. Kelder. Estimates of lightning NOx production from GOME satellite observations. *Atmospheric Chemistry and Physics Discussions*, 5(3) :3047–3104, 2005.

C. Bohren and D. Huffman. *Absorption and Scattering of Light by Small Particles*. J. Wiley and Sons, 1983.

J.-J. Boreux, P. Naveau, O. Guin, L. Perreault, and J. Bernier. Extracting a common high frequency signal from northern quebec black spruce tree-rings with a bayesian hierarchical model. *Climate of the Past*, 5(4) :607–613, 2009. doi : 10.5194/cp-5-607-2009.

R. Bourbonnais and M. Terraza. *Analyse des séries temporelles*. 2004. ISBN 2100484362.

R. Bradley. The explosive volcanic eruption signal in Northern Hemisphere continental temperature records. *Climatic Change*, 12 :221–243, 1988.

K. Briffa, P. Jones, F. Schweingruber, and T. Osborn. Influence of volcanic eruptions on Northern Hemisphere summer temperature over the past 600 years. *Nature*, 393 : 450–455, 1998.

P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer, Germany, Heidelberg, 2002. ISBN 978-0-387-95351-9.

E. Castellano, S. Becagli, M. Hannson, M. Hutterli, J. R. Petit, M. Rampino, M. Severi, J. P. Steffensen, R. Traversi, and R. Udisti. *JGR*, 110, 2005. doi : 10.1029/2004JD005259.

E. Castellano, S. Becagli, J. Jouzel, A. Migliori, M. Severi, J. Steffensen, R. Traversi, and R. Udisti. Volcanic eruption frequency over the last 45 ky as recorded in Epica-Dome C ice core (East Antarctica) and its relationship with climatic changes. *Global and Planetary Change*, 42(1-4) :195–205, 2004. doi : {10.1016/j.gloplacha.2003.11.007}.

H. Caussinus and O. Mestre. Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society*, 53 :405–425, 2004. doi : 10.1111/j.1467-9876.2004.05155.x.

S. I. Chang and K. Zhang. Statistical process control for variance shift detections of multivariate autocorrelated processes. *Quality Technology and Quantitative Management*, 4 :413–435, 2007.

P. Chazette, C. David, L. J, G. S, and M. G. Comparative lidar study of the optical, geometrical and dynamical properties of stratospheric post-volcanic aerosols, following the eruptions of el chichon and mount pinatubo. *Journal of geophysical research*, 100 : 195–207, 1995.

P. Chazette, J. Pelon, and G. Megie. Determination by spaceborne backscatter lidar of the structural parameters of atmospheric scattering layer. *Applied Optics*, 40(21) :3428–3440, 2001. doi : 10.1364/A0.40.003428.

S. Chib. Estimation and comparison of multiple change-point models. *J.Econometrics*, 86 :221–241, 1998. doi : 10.1016/S0304-4076(97)00115-2.

L. Cirbus Sloan and D. Pollard. Polar stratospheric clouds : A high latitude warming mechanism in an ancient greenhouse world. *Geophysical Research Letters*, 25, 1998.

H. Clausen, C. Hammer, C. Hvidberg, D. Dahl-Jensen, J. Steffensen, J. Kipfstuhl, and M. Legrand. A comparison of volcanic records over the past 4000 years from the greenland ice core project and dye 3 greenland ice cores. *JGR*, 102, 1997.

R. Collis and P. Russell. Lidar measurement of particles and gases by elastic backscattering and differential absorption, laser monitoring of the atmosphere. *Applied Physics : Laser Monitoring of the Atmosphere*, 14, 1976.

T. Crowley. Causes of climate change over the past 1000 years. *Science*, 289 :270–277, 2000.

T. J. Crowley and K.-Y. Kim. Modeling the temperature response to forced climate change over the last six centuries. *Geophys. Res. Lett.*, 26, 1999.

R. Dalang and D. Conus. *Introduction à la théorie des probabilités*. 2008.

C. David. *Etude des nuages stratosphériques polaires et des aérosols volcaniques en régions polaires par sondage laser*. PhD thesis, ParisVI University, 1995.

C. David, S. Bekki, S. Godin, G. Mégie, and M. P. Chipperfield. Polar stratospheric clouds climatology over dumont d'urville between 1989 and 1993 and the influence of volcanic aerosols on their formation. *Journal of geophysical research*, pages 163–180, 1998.

C. David, S. Bekki, N. Berdunov, M. Marchand, M. Snels, and G. Mégie. Classification and scales of antarctic polar stratospheric clouds using wavelet decomposition. *Journal of Atmospheric and Solar-terrestrial physics*, 67 :293–300, 2004.

C. David, P. Keckhut, A. Armetta, J. Jumelet, M. Marchand, and S. Bekki. Radiosondes stratospheric temperatures from 1957 to 2008 at dumont d'urville (antarctica) : trends and links with polar stratospheric clouds. *Atmospheric Chemistry and Physics Discussions*, 9 :25687–25722, 2009.

R. A. Davis, T. C. Lee, and G. A. Rodriguez-Yam. Structural break estimation for nonstationary time series signals. *J.American Statist. Assoc.101*, pages 229–239, 2006.

L. De Montera. *Etude de la variabilité micro-échelle des précipitations : Application à la propagation des ondes millimétriques en SATCOM*. PhD thesis, Université de Versailles-Saint Quentin en Yvelines, 2008.

R. Delmas, G. Mégie, and V.-H. Peuch. *Physique et chimie de l'atmosphère*. Belin, France, Paris, 2005.

A. Dempster, N. Laird, D. B. Rubin, L. Cirbus Sloan, and D. Pollard. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39, 1977.

J. Diard, P. Bessiere, and E. Mazer. A Survey of Probabilistic Models Using the Bayesian Programming Methodology as a Unifying Framework. In *International Conference on Computational Intelligence, Robotics and Autonomous Systems (IEEE-CIRAS), Singapore –*, France, 2003.

A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

J.-J. Droesbeke, M. Lejeune, and G. Saporta. *Méthodes statistique pour données qualitatives*. Technip, France, Paris, 2005. ISBN 2-7108-0855-2.

A. Dubietis, P. Dalin, R. Balciunas, and K. Cernis. Observations of noctilucent clouds from lithuania. *Journal of Atmospheric and Solar-Terrestrial Physics*, 72(14-15) :1090 – 1099, 2010. ISSN 1364-6826. doi : DOI:10.1016/j.jastp.2010.07.004.

H. Egon and P. Porée. *Statistique et probabilités en production industrielle : Volume 2, Contrôle et maîtrise de la qualité, fiabilité, problèmes et exercices corrigés*. Hermann, éditeurs des sciences et des arts, France, Paris, 2003.

A. Einstein. Uber die von der molekularkinetischen theorie der warme geforderte bewegung von in ruhenden flussigkeiten suspendierten teilchen. *Annalen der Physics*, 322 : 549–560, 1905.

G. Evensen. The ensemble kalman filter : theoretical formulation and practical implementation. *Ocean Dynamics*, 53, 2003. ISSN 1616-7341.

G. Evensen. *Data Assimilation : The Ensemble Kalman Filter*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 354038300X.

G. Evensen. *Data Assimilation : The Ensemble Kalman Filter, 2nd Edition*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2009. ISBN 978-3-642-03710-8.

F. Fernald, B. Herman, and J. Reagan. Determination of aerosol height distributions by lidar. *J. Appl. Meteorol.*, 11 :482, 1972.

F. Fierli, A. Hauchecorne, and B. Knudsen. Analysis of polar stratospheric clouds using temperature and aerosols measured by alomar r/m/r lidar. *Journal of geophysical research*, 106 :127, 2001.

G. Fiocco and L. Smullins. Detection of scattering layers in the upper atmosphere (60–140 km) by optical radar. *Nature*, 199 :1275, 1963.

J. Fisher-Box. Guinness, Gosset, Fisher and Small Samples. *Statistical Science*, 2 :45–52, 1987.

B. Fontaine and S. Louvet. Sudan-Sahel rainfallonset : Definition of an objective index, types of years, and experimental hindcasts. *J. Geophys. Res.*, page 111, 2006.

B. Fontaine, S. Louvet, and P. Roucou. Definition and predictability of an OLR-based West African monsoon onset. *International Journal of Climatology*, 28 :1787–1798, 2008.

A. Fraser, F. Goutail, C. A. Mclinden, S. M. L. Melo, and K. Strong. Lightning-produced NO2 observed by two ground-based UV-visible spectrometers at Vanscoy, Saskatchewan in August 2004. *Atmospheric Chemistry and Physics*, 7(6) :1692, 2007.

S. Frontier, D. Davoult, V. Gentilhomme, and Y. Lagadeux. *Statistique pour les sciences de la vie et de l'environnement*. Dunod, Science Sup, France, Paris, 2007.

C. Gao, A. Robock, S. Self, J. B. Witter, J. Steffenson, H. B. Clausen, M. Siggaard-Andersen, S. Johnsen, P. A. Mayewski, and C. Ammann. The 1452 or 1453 a.d. kuwae eruption signal derived from multiple ice core records : Greatest volcanic sulfate event of the past 700 years. *JGR*, 111, 2006. doi : 10.1029/2005JD006710.

C. Gao, L. Oman, A. Robock, and G. Stenchikov. Atmospheric volcanic loading derived from bipolar ice cores : Accounting for the spatial distribution of volcanic deposition. *J. Geophys. Res.*, 112, 2007. doi : 10.1029/2006JD007461.

C. Gao, A. Robock, and C. Ammann. Volcanic forcing of climate over the past 1500 years : An improved ice-core based index for climate models. *Journal of Geophysical Research*, 2008.

J. Gazeaux, D. Batista, C. Ammann, P. Naveau, C. Jégat, and G. C. Extracting common pulse-like signals from multiple ice core time series. *Submitted to Journal of Geophys. Res.*, 2009.

B. Goldfarb and C. Pardoux. *Introduction à la méthode statistique*. Dunod, Science Sup, France, Paris, 2007.

S. Gosset, Willima. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

F. N. Gumedze, S. J. Welham, B. J. Gogel, and R. Thompson. A variance shift model for detection of outliers in the linear mixed model. *Comput. Stat. Data Anal.*, 54 :2128–2144, September 2010. ISSN 0167-9473. doi : http://dx.doi.org/10.1016/j.csda.2010.03.019.

W. Guo, Y. Wang, and M. B. Brown. A signal extraction approach to modeling hormone time series with pulses and a changing baseline. *Journal of the American Statistical Association*, 94 :746–756, 1998.

W. Guo, Y. Wang, and M. Brown. A signal extraction approach to modeling hormone time series with pulses and a changing baseline. *J. of Am. Stat. Ass.*, 94 :746–756, 1999.

Y. Guo and Y. Ding. Long-term free-atmosphere temperature trends in china derived from homogenized in situ radiosonde temperature series. *Journal of Climate*, 22(4) : 1037–1051, 2009. doi : 10.1175/2008JCLI2480.1.

S. Hagos and K. Cook. Dynamics of the West African monsoon jump. *Journal of Climate*, 20 :5264–5284, 2007.

T. M. Hamill and C. Snyder. A hybrid ensemble kalman filter / 3d-variational analysis scheme. *American Meteorological Society*, 2000.

C. Hammer. Past volcanism revealed by Greenland Ice Sheet impurities. *Nature*, 270 : 482–486, 1977.

A. Hannart and P. Naveau. Bayesian multiple change points and segmentation : Application to homogenization of climatic series. *Water Resour. Res.*, 45, 2009. doi : 10.1029/2008WR007689.

J. Hansen and Coauthors. Climate forcings in goddard institute for space studies si2000 simulations. *J Geophys Res-Atmos*, 107, 2002.

G. Hegerl, T.Crowley, S. Baum, K. Kim, and W. Hyde. Detection of volcanic, solar and greenhouse gas signals in paleo-reconstructions of nothern hemispheric temperature. *Geophys. Res. Lett.*, 30(5) :1242, 2003. doi : 10.1029/2002GL016271.

G. Hegerl, T. Crowley, W. Hyde, and D. Frame. Climate sensitivity constrained by temperature reconstructions over the past seven centuries. *Nature*, 440 :1029–1032, 2006. doi : 10.1038/nature04679.

G. C. Hegerl, F. W. Zwiers, P. Braconnot, N. P. Gillett, Y. Luo, J. A. M. Orsini, N. Nicholls, J. E. Penner, and P. A. Stott. 9 supplementary materials understanding and attributing climate change coordinating lead authors :, 2007. Contribution of Working Group I to the fourth assessment report of the Intergovernmental Panel on climate change.

I. B. Hossack, J. H. Pollard, and B. Zehnwirth. *Introductory statistics with applications in general insurance*. Cambridge University Press, 1999. ISBN 0 521 65534 X.

M. K. Hughes and C. M. Ammann. The future of the past - an earth system framework for high resolution paleoclimatology : Editorial essay. *Climatic change*, 94, 2009. doi : 10.1007/s10584-009-9588-0.

S. Janicot, C. D. Thorncroft, Alia, and al. Large-scale overview of the summer monsoon over West africa during the AMMA field experiment in 2006. *Annales Geophysicae*, 26 :2569–2595, 2008.

P. Jones, A. Moberg, T. Osborn, and K. Briffa. Surface climate responses to explosive volcanic eruptions seen in long european temperature records and mid-to-high latitude tree-ring density around the northern hemisphere. *Geophysical monograph*, 139 :239–254, 2003. ISSN 0065-8448.

C. E. Junge, C. W. Chagnon, and J. E. Manson. A world-wide stratospheric aerosol layer. *Science*, 133 :1478–1479, 1961.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82 :35–45, 1960.

R. E. Kalman and R. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83 :95–107, 1961.

M. Kanamitsu, W. Ebisuzaki, J. Woollen, S. Yang, J. Hnilo, M. Fiorino, and G. Potter. NCEP-DOE AMIP-II reanalysis (R-2). *Bulletin of the American Meteorological Society*, 83 :1631–1643, 2002.

R. W. Katz. Sir gilbert walker and a connection between el niño and statistics. *Statistical Science*, 17, 2002. doi : 10.1214/ss/1023799000.

R. Keen. Volcanic aerosol optical thicknesses since 1960. *Global Volcanism Network*, 26, 2001.

J. Klett. Stable analytical inversion solution for processing lidar returns. *Appl. Opt.*, 20 : 2, 1981.

J. Klett. Lidar inversion with variable backscatter/extinction ratios. *Appl. Opt.*, 24 :1638, 1985.

R. Kohn and C. F. Ansley. Signal extraction for finite nonstationary time series. *Biometrika*, 74 :411–421, 1987.

V. A. Kovalev and W. E. Eichinger. *Elastic Lidar : Theory, Practice, and Analysis Methods*. Wiley, USA,Hoboken, NJ, 2004. ISBN 978-0-471-20171-7.

A. Kurbatov, G. Zielinski, N. Dunbar, P. Mayewski, and S. Meyerson, E.and Sneed. A 12,000 year record of explosive volcanism in the siple dome ice core, west antarctica. *Journal of Geophysical Research*, 111, 2006. doi : 10.1029/2005JD006072.

T. Lachlan-Cope, W. Connolley, J. Turner, H. Roscoe, G. Marshall, S. Colwell, M. Hopfner, and W. Ingram. Antarctic winter tropospheric warming - the potential role of polar stratospheric clouds, a sensity study. *Atmopsheric Sceince Letters*, 10, 2009. doi : 10.1002/asl.237.

H. Lamb. Volcanic dust in the atmosphere ; with a chronology and assessment of its meteorological significance. *Transactions of the Royal Philosophical Society of London*, A266 :425–533, 1970.

C. J. Langway, K. Osada, H. Clausen, C. U. Hammer, and H. Shoji. A 10-century comparison of prominent bipolar volcanic events in ice cores. *JGR*, 100, 1995.

M. Lavielle and E. LeBarbier. "an application of MCMC methods for the multiple change-points problem". *Signal Process*, 81 :39–53, 2001. doi : 10.1016/S0165-1684(00) 00189-4.

L. Le Barbé, T. Lebel, and D. Tapsoba. Rainfall variability in west africa during the years 1950-90. *Journal of Climate*, 15(2) :187–202, 2002. doi : 10.1175/1520-0442(2002) 0152.0.CO;2.

B. Legras, O. Mestre, E. Bard, and P. Yiou. On misleading solar-climate relationship. *Climate of the Past Discussions*, 6 :767–800, 2010. doi : 10.5194/cpd-6-767-2010.

B. Liebmann and C. Smith. Description of a complete outgoing longwave radiation dataset. *Bull. Amer. Meteor. Soc.*, 77 :1275–1277, 1996.

N. Mahowald, C. Luo, J. del Corral, and C. Zender. Interannual variability in atmospheric mineral aerosols from a 22-year model simulation and observational data. *Journal Of Geophysical Research-Atmosphere*, 108(D12), 2003. doi : {10.1029/2002JD002821}.

S. Malardel. *Fondamentaux de Météorologie*. Cépadués, France, Toulouse, 2005.

E. D. Maloney and J. Shaman. Intraseasonal variability of the West African monsoon and Atlantic ITCZ. *Journal Of Climate*, 21(12) :2898–2918, 2008. doi : {10.1175/ 2007JCLI1999.1}.

R. Mankiewicz. *Histoire des mathématiques*. Seuil, France,Paris, 2001. ISBN 2-02-055073-3.

M. Manoliu and T. Stathis. Energy futures prices : Term structure models with kalman filter estimation. *MSIS Department and Center for Computational Finance, University of Texas at Austin*, 2002.

J. C. Maxwell. On the dynamical theory of gases. *Phil. Mag.*, 32 :390–393, 1866.

P. Mayewski, W. Lyons, M. Spencer, M. Twickler, P. Koci, P. Dansgaard, C. Davidson, and R. Honrath. Sulfate and Nitrate concentrations from a South Greenland ice core. *Science*, 232 :975–977, 1986.

R. E. McCulloch and R. S. Tsay. Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association*, 88 : 968–978, 1993. ISSN 01621459.

A. McCutcheon. *Latent class analysis*. Sage, USA, Beverly Hills, 1987.

R. J. Meinhold and N. D. Singpurwalla. Understanding the Kalman Filter. *The American Statistician*, 37 :123–127, 1983.

O. Mestre and H. Caussinus. A correction model for homogenisation of long instrumental data series. In India, MB and Bonillo, DL, editor, *Detecting and modelling regional climate change*, pages 13–19, Berlain, Germany, 2001. Springer-Verlag Berlin. ISBN 3-540-42239-0.

J. M. Moisselin, M. Schneider, C. Canellas, and O. Mestre. Changements climatiques en france au 20ème siècle - Étude des longues séries de données homogénéisées françaises de précipitations et températures. *La Météorologie*, 38 :45–56, 2002.

A. M. Mood. *Introduction to the Theory of Statistics (Third Edition)*. McGraw-Hill, 1974. ISBN 0-07-042864-6.

Y. Morille, M. Haeffelin, P. Drobinski, and J. Pelon. Strat : An automated algorithm to retrieve the vertical structure of the atmosphere from single-channel lidar data. *Journal of Atmospheric and Oceanic Technology*, 24(5) :761–775, 2007. doi : 10.1175/JTECH2008.1.

E. Mosley-Thompson, L. Thompson, J. Dai, M. Davis, and P. Lin. Climate of the last 500 years : High resolution ice core records. *Quat. Sci. Rev.*, 12, 1993.

P. Naveau and C. Ammann. Statistical distributions of ice core sulfate from climatically relevant volcanic eruptions. *Geophysical Research Letters*, 32, 2005. doi : 10.1029/2004GL021732.

P. Naveau, C. Ammann, H. Oh, and W. Guo. An automatic statistical methodology to extract pulse like forcing factors in climatic time series : Application to volcanic events. *Geophysical monograph*, 139 :177–186, 2003. ISSN 0065-8448.

A. Neftel, J. Beer, H. Oeschger, F. ZÃ$\frac{1}{4}$rcher, and R. Finkel. Sulphate and nitrate concentrations in snow from South Greenland 1895-1978. *Nature*, 314 :611–613, 1985.

C. Newhall and S. Self. The volcanic explosivity indes (VEI) : An estimate of explosive magnitude for historical volcanism. *Journal of Geophysical Research*, 87 :1231–1238, 1982.

S. Nicholson. Rainfall and atmospheric circulation during drought periods and wetter years in West Africa. *Mon. Wea. Rev.*, 109 :2191–2208, 1981.

Ouachani, Bargaoui, and Ouarda. *Integration of a Kalman filter in the HBV hydrological model for runoff forecasting*. Taylor et Francis, GB, London, 2010. ISBN 0262-6667.

J. Palais, M. Germani, and G. Zielinski. Interhemispheric transport of volcanic ash from a 1259 ad volcanic-eruption to the greenland and antarctic ice sheets. *GRL*, 19, 1992.

T. Palmer, A. S.and van Ommen, M. A. J. Curran, V. Morgan, J. Souney, and P. Mayewski. High-precision dating of volcanic events (a.d. 1301-1995) using ice cores from law dome, antarctica. *JGR*, 106, 2001.

J. Pappel. Lidar : Range-resolved optical remote sensing of the atmosphere,. *Optik - International Journal for Light and Electron Optics*, 117(7) :308 – 308, 2006. ISSN 0030-4026. doi : DOI:10.1016/j.ijleo.2006.03.006.

E. Parzen. On estimation of a probability density function and mode. *Ann Math Statist*, 33 :1065–1076, 1962.

J. Perrin. *Les atomes*. Félix Alcan, 1913.

T. Peter. Microphysics and heterogeneous chemistry of polar stratospheric clouds. *Annu. Rev. Phys. Chem*, 48 :785–822, 1997.

C. A. Pires, O. Talagrand, and M. Bocquet. Diagnosis and impacts of non-Gaussianity of innovations in data assimilation. *Physica D-Nonlinear Phenomena*, 239(17) :1701–1717, 2010. ISSN 0167-2789. doi : {10.1016/j.physd.2010.05.006}.

C. M. R. Platt. Remote sounding of high clouds : I. calculation of visible and infrared optical properties from lidar and radiometer measurements. *Journal of Applied Meteorology*, 18(9) :1130–1143, 1979. doi : 10.1175/1520-0450(1979)018.

J.-L. Puget and R. Blanchete. Le changement climatique. Technical report, Académie des Sciences, France, Paris, 2010.

R Development Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

J. Reeves, J. Chen, X. L. Wang, R. Lund, and Q. Q. Lu. A review and comparison of change-point detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46 :900–915, 2007. doi : 10.1175/JAM2493.1.

A. Ribes. *Détection statistique de changement climatique*. PhD thesis, Université Toulouse III - Paul Sabatier, 2009.

A. Ribes, J.-M. Azais, and S. Planton. A method for regional climate change detection using smooth temporal patterns. *Climate Dynamics*, 35 :391–406, 2010. doi : 10.1007/s00382-009-0670-0.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(2) :445–471, 1978.

A. Robock. Volcanic eruptions and climate. *Reviews of Geophysics*, 38 :191–219, 2000.

A. Robock and M. Free. Ice cores as an index of global volcanism from 1850 to the present. *Journal of Geophysical Research*, 100 :549–11, 1995.

A. Robock and J.Mao. The volcanic signal in surface temperature observations. *Journal of Climate*, 8 :1086–1103, 1995.

J. Rosen, D. Hofmann, and J. Laby. Stratospheric aerosol measurements ii : The worldwide distribution. *J. Atm. Sc.*, 32 :1457–1462, 1975.

Saporta and CoAuthors. Modèles à variables latentes et modèles de mélange. pages 243–284. CIRM, 2008.

S. Sarkka, V. Aki, and J. Lampinen. Time series prediction by kalman smoother with cross-validated noise density, 2006.

M. Sato, J. Hansen, M. P. McCormick, and J. Pollack. Stratospheric aerosol optical depth, 1850-1990. *Journal of geophysical research*, 98, 1993.

G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6, 1978. doi : 10.1214/aos/1176344136.

N. Shepard. Partially non-gaussian state-space models. *Biometrika*, 81 :115–131, 1994.

S. T. Shipley, D. H. Tracy, E. W. Eloranta, J. T. Trauger, J. T. Sroga, F. L. Roesler, and J. A. Weinman. High spectral resolution lidar to measure optical scattering properties of atmospheric aerosols. 1 : Theory and instrumentation. *Appl. Opt.*, 22(23) :3716–3724, 1983. doi : 10.1364/AO.22.003716.

T. Simkin and L. Siebert. *Volcanoes of the World*, volume 2 ed. Geoscience Press, Inc. and Smithsonian Institution, Tucson, AR, 1994.

M. Sing Wong, J. E. Nichol, and K. Ho Lee. Modeling of aerosol vertical profiles using gis and remote sensing. *Sensors*, 9 :4380–4389, 2009.

SPARC. *SPARC CCMVal (2010), SPARC Report on the Evaluation of Chemistry-Climate Models*. http ://www.atmosp.physics.utoronto.ca/SPARC, 2010. SPARC Report No. 5.

R. C. Sprinthall. *Basic Statistical Analysis*. Prentice Hall, USA, New Jersey, 2009.

J. Staehelin, C. Vogler, and S. Brannimann. The long history of ozone measurements : Climatological information derived from long ozone records. In C. Zerefos, G. Contopoulos, and G. Skalkeas, editors, *Twenty Years of Ozone Decline*, pages 119–131. Springer Netherlands, 2009. ISBN 978-90-481-2469-5.

H. Stark and J. W. Woods. *Probability and Random Processes with Applications to Signal Processing*, volume 3 ed. Prentice Hall, USA, New Jersey, 2002.

R. Stothers. The great dry fog of 1783. *Climatic Change*, 32, 1996a.

R. Stothers. Major optical depth perturbations to the stratosphere from volcanic eruptions : Pyrheliometric period, 1881-1960. *Journal of Geophysical research*, 101, 1996b.

B. Sultan and S. Janicot. Abrupt shift of the ITCZ over West Africa and intra-seasonal variability. *Geophys. Res. Lett.*, 27 :3353–3356, 2000.

B. Sultan and S. Janicot. The west african monsoon dynamics. part II : The preonset and onset of the summer monsoon. *Journal of Climate*, 16(21) :3407–3427, 2003. doi : 10.1175/1520-0442(2003)0162.0.CO;2.

A. Tabazadeh, R. Turco, and M. Jacobson. A model for studying the composition and chemical effects of stratospheric aerosols. *J. Geophys. Res.*, 99 :12897–12914, 1994.

O. Talagrand. A posteriori validation of assimilation algorithms. In Swinbank, R and Shutyaev, V and Lahoz, WA, editor, *Data Assimilation for The Earth System*, volume 26 of *NATO Science Series IV Earth and Environmental Sciences*, pages 85–95. SpringerR, 2003. ISBN 1-4020-1592-5.

R. Taton. *Histoire générale des sciences : La science contemporaine - Le XXe siècle*. Presse Universitaire de France, France,Paris, 1995. ISBN 2-13-046889-6.

T. Thordarson and S. Self. Atmospheric and environmental effects of the 1783–1784 laki eruption : A review and reassessment. *JGR*, 108, 2003. doi : 10.1029/2001JD002042.

C. Thorncroft and K. Hodges. African easterly wave variability and its relationship to Atlantic tropical cyclone activity. *Journal Of Climate*, 14(6) :1166–1179, 2001. ISSN 0894-8755.

S. Tufféry. *Data mining et statistique décisionnelle*. Technip, France, Paris, 2010. ISBN 978271080946-3.

Villien. *Prévision de trajectoires 3-D en temps réel*. PhD thesis, Université Louis Pasteur Strasbourg I, 2006.

C. Voigt, J. Schreiner, A. Kohlmann, P. Zink, K. Mauersberger, N. Larsen, T. Deshler, C. Kröger, A. Rosen, J.and Adriani, G. Cairo, F.and Donfrancesco, M. Viterbini, J. Ovarlez, C. Ovarlez, H.and David, and A. Dörnbrack. Nitric acid trihydrate (nat) in polar stratospheric clouds. *Science*, 290 :1756–1758, 2000.

G. Von Cossart, P. Hoffmann, U. vonZahn, P. Keckhut, and A. Hauchecorne. Mid-latitude noctilucent cloud observations by lidar. *Geophysical Research Letters*, 23(21) :2919–2922, 1996.

H. Von Storch and F. W. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, 1999. ISBN 0-521-45071-3.

U. Von Zahn, G. Von Cossart, J. Fiedler, K. H. Fricke, G. Nelke, G. Baumgarten, D. Rees, A. Hauchecorne, and K. Adolfsen. The alomar rayleigh/mie/raman lidar :objectives,configuration and performance. *Annales Geophysicae*, 18 :815–833, 2000.

G. Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. of the Royal Stat. Soc.*, 40 :364, 1978.

S. Wastegard and S. M. Davies. An overview of distal tephrochronology in northern Europe during the last 1000 years. *Journal of Quaternery Science*, 24 :500–512, 2009. doi : {10.1002/jqs.1269}.

W. E. Wecker and C. F. Ansley. The signal extraction approach to nonlinear regression and spline smoothing. *J. of the Amer.Stat. Ass.*, 78 :81–89, 1983.

G. Welch and G. Bishop. An introduction to the Kalman Filter. Technical report, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1995.

G. Welch and G. Bishop. An introduction to the Kalman Filter. Technical report, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 2006.

T. M. L. Wigley, C. Ammann, B. D. Santer, and S. Raper. Effect of climate sensitivity on the response to volcanic forcing. *Journal of Geophysical Research*, 110, 2005. doi : 10.1029/2004JD005557.

WMO. Scientific assessment of ozone depletion :2006, global ozone research and monitoring project, 2007.

J. Woods. Two-dimensional kalman filtering. In *Two-Dimensional Digital Signal Processing I*, volume 42, pages 155–205. Springer Berlin / Heidelberg, 1981.

G. Zielinski. Use of paleo-records in determining variability within the volcanism-climate system. *Quaternary Science Reviews*, 19 :417–438, 2000.

G. Zielinski, P. Mayewski, L. Meeker, S. Whitlow, M. Twickler, M. Morrison, D. Meese, A. Gow, and R. Alley. Record of volcanism since 7000 B.C. from the GISP2 Greenland ice core and implications for the volcano-climate system. *Science*, 264 :948–952, 1994.

F. Zwiers and H. Von Storch. On the role of statistics in climate research. *International Journal Of Climatology*, 24(6) :665–680, 2004. ISSN 0899-8418.

Appendices

Annexe A

Rappel de définitions et notions de probabilités : Théorie et Méthodologie

Ce travail de thèse se focalise sur la recherche, l'extraction et la caractérisation d'évènements cachés dans des séries de données issues de problèmes rencontrés en géophysique. Il est présenté dans ce chapitre les différentes notions et outils de base utilisés au cours des différentes études. Nous présentons tout d'abord quelques définitions et notions élémentaires de probabilité qui seront nécessaires pour la compréhension des différents développements suivants. Pour plus de détails sur ces notions nous conseillons l'ouvrage [Dalang and Conus, 2008].

1 étude de variables aléatoires continues : Densité de probabilité

De manière informelle, les probabilités sont étudiées généralement par l'intermédiaire de la fonction de répartition, qui caractérise la fonction de probabilité, et qui a pour avantage principal que son étude ne nécessite pas de différencier les cas de probabilités discrètes des cas de probabilités continues. La fonction de répartition d'une variable aléatoire définit le poids cumulé compris entre 0 et 1 porté par un ensemble croissant des réalisations possibles de la variable aléatoire. Cette thèse est essentiellement axée sur l'étude de probabilités continues et ainsi se concentre davantage sur l'étude de la densité de probabilité plutôt que sur la fonction de répartition. Cette section introduit le champ des probabilités continues et explique les différents développements faits au cours de la thèse.

1.1 Définir et caractériser une variable aléatoire

De manière plus formelle maintenant, une variable aléatoire définie sur une espace mesurable (Ω,$\mathcal{F}$) est une application de Ω dans $\mathbb{R}$ vérifiant :

$$\forall x \in \mathbb{R}, \quad X^{-1}(]-\infty, x[) = \{\omega \in \Omega, X(\omega) \leq x\} \in \mathcal{F} \tag{A.1}$$

En d'autres termes, il s'agit de toute application mesurable de Ω vers $\mathbb{R}$ munis de leurs tribues respectives. En utilisant la notion de tribu borélienne définie sur $\mathbb{R}$ (la plus petite tribu qui contient les intervalles de $\mathbb{R}$), une variable aléatoire X vérifie :

$$\forall B \in \mathcal{B}, \quad X^{-1}(B) = \{\omega \in \Omega, X(\omega \in B)\} \in \mathcal{F} \tag{A.2}$$

L'ensemble $X^{-1}(B)$ ainsi défini est noté généralement $\{X \in B\}$, de manière similaire $X^{-1}(]-\infty, x[)$ est noté $\{X \leq x\}$. Les ensembles sont plus communément appelés évènements aléatoires.

1.2 Densité de probabilité

Agrémenté d'une mesure de probabilité $\mathbb{P}$, l'espace mesurable précédent devient espace de probabilité (Ω,$\mathcal{F}$,$\mathbb{P}$). Sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, la mesure de probabilité $\mathbb{P}$, également noté $\mathbb{P}_X$ pour signifier qu'il s'agit de la mesure de la variable aléatoire X, est définie par :

$$\mathbb{P}_X : \mathcal{B}(\mathbb{R}) \to [0,1], \quad \mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) \tag{A.3}$$

La loi $\mathbb{P}_X$ de X est bien une mesure, dans le sens où, $\mathbb{P}(X^{-1}(\mathbb{R})) = \mathbb{P}(\Omega) = 1, \mathbb{P}(X^{-1}(\emptyset)) = \mathbb{P}(\emptyset) = 0$, et $\mathbb{P}_X$ est σ-additive, c'est à dire que la mesure de la réunion dénombrable de parties deux à deux disjointes est égale à la somme des mesures des parties.

Cette mesure de probabilité nous permet maintenant de définir la densité de probabilité p d'une variable aléatoire continue X sur $\mathbb{R}$:

$$\forall x \in \mathbb{R}, \quad F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} p(t)dt \tag{A.4}$$

Avec $p : \mathbb{R} \to \mathbb{R}$, positive, Lebesgue-intégrable et telle que $\int_{\mathbb{R}} p(t)dt = 1$. On dit alors

que X a pour densité p. Et on définit la probabilité d'un évènement $B \in \mathcal{B}(\mathbb{R})$par :

$$\mathbb{P}(B) = \int_B p(t)dt \tag{A.5}$$

On finit cette sous section en donnant quelques exemples de densités de probabilité de variables aléatoires continues définies sur $\mathbb{R}$:

La loi Uniforme : $\mathcal{U}$, $\forall x \in [a, b[\ p(x) = (b-a)^{-1}\mathbf{1}_{\mathbf{[a,b[}}$
La loi exponentielle : $\mathcal{E}$, $\forall x \in \mathbb{R},\ \lambda > 0,\ p(x) = \lambda exp(-\lambda x)\mathbf{1}_{\mathbf{x>0}}$
La loi Gaussienne : $\mathcal{N}$, $\forall x \in \mathbb{R},\ (\mu, \sigma) \in \mathbb{R}^2,\ p(x) = (2\pi\sigma^2)^{-1/2}exp(-(x-\mu)/(2\sigma^2))$
Il est à noter que les densités de probabilité citées ci-dessus mettent en jeux un certain nombre de paramètres : (a, b) pour $\mathcal{U}$, (λ) pour $\mathcal{E}$, ou encore (μ, σ) pour $\mathcal{N}$.

On appelle fonction de vraisemblance, une densité de probabilité qu'on ne considère plus comme fonction de x, mais comme fonction de ses paramètres caractéristiques. Les fonctions de vraisemblance, généralement notés $\mathcal{L}$, des densités précédentes sont données par :
La loi Uniforme : $\mathcal{L}_{\mathcal{U}}(a, b) = (b-a)^{-1}\mathbf{1}_{\mathbf{[a,b[}}$
La loi exponentielle : $\mathcal{L}_{\mathcal{E}}(\lambda) = \lambda exp(-\lambda x)\mathbf{1}_{\mathbf{x>0}}$
La loi Gaussienne : $\mathcal{L}_{\mathcal{N}}(\mu, \sigma) = (2\pi\sigma^2)^{-1/2}exp(-(x-\mu)/(2\sigma^2))$

1.3 Caractérisation de variables aléatoires par les moments

Il est parfois plus aisé et plus efficace de caractériser une variable aléatoire, ou un ensemble de variables aléatoires non par leur densité mais par un nombre restreint d'information telle que la valeur moyenne (aussi appelé moment d'ordre un) ou la dispersion de la variable aléatoire autour de la valeur moyenne (appelé moment d'ordre deux). Ces moments sont donnés par :

$$\mathbb{E}(X) = \int_{\mathbb{R}} tp(t)dt \tag{A.6}$$

$$\mathbb{V}(X) = \int_{\mathbb{R}} t^2 p(t)dt - \left(\int_{\mathbb{R}} tp(t)dt\right)^2 \tag{A.7}$$

Certaines densités de probabilité sont entièrement définies par un ou deux de ces moments. C'est le cas pour les trois exemples de densité cités précédemment. En effet, la valeur moyenne d'une loi de type $\mathcal{U}(a,b)$ est donnée par $\mathbb{E}_{\mathcal{U}} = \frac{a+b}{2}$ et sa variance par $\mathbb{V}_{\mathcal{U}} = \frac{(b-a)^2}{12}$. La connaissance de $\mathbb{E}_{\mathcal{U}}$ et $\mathbb{V}_{\mathcal{U}}$ permettant de calculer (a,b) donc de retrouver la densité initiale.
Dans certains cas, en revanche, caractériser des densités de probabilités nécessite la définition de moments d'ordre supérieur à deux :

$$\mathbb{E}(X^k) = \int_{\mathbb{R}} t^k p(t) dt \tag{A.8}$$

Dans le chapitre 3 de la thèse nous nous intéresserons particulièrement à la manipulation de la fonction de vraisemblance dans le cas de variables Gaussiennes, et plus particulièrement au calcul du maximum du rapport de vraisemblance entre deux hypothèses qui nous permettra de définir les paramètres d'une distribution. Dans les parties 4 et 5, nous nous intéresserons à définir les moments d'ordre un et deux de densité des probabilités.

2 Les processus stochastiques

Un processus stochastique, signal aléatoire indexé par le temps et prenant ses valeurs dans un espace continu est noté :

$$\{x_t(\omega) | t \in T\} \tag{A.9}$$

$x_t(\omega)$ est une fonction de deux paramètres : le temps t et ω, un paramètre lié au résultat d'une expérience aléatoire. Un tel processus est une variable aléatoire qui varie en fonction du temps, dont à chaque instant, la réalisation est le résultat d'une expérience aléatoire imprévisible. Il peut s'agir par exemple de la vitesse du vent sur une station, ou encore la concentration d'un gaz dans l'atmosphère, ou encore la hauteur des vagues mesures par houlographes.

Par définition, un processus stochastique n'est pas parfaitement prévisible, tout comme l'est la réalisation d'une variable aléatoire. Certaines caractéristiques permettent de les étudier, telles que la stationnarité, l'hétéroscedasticité ou encore l'ergodisme présentées ci dessous.

Soit x_t un processus stochastique ayant pour fonction de densité de probabilité conjointe $f_{x_{t_1},..,x_{t_n}}(\alpha_1,..,\alpha_k)$, où $(x_{t_1},..,x_{t_n})$ représente l'indice des temps et $(\alpha_1,..,\alpha_k)$ les para-

mètres caractérisant la densité f. Alors, un processus stochastique est dit stationnaire au sens strict sur un intervalle I, si pour toute partition $t_1, .., t_n$ de I :

$$\forall n, \forall \theta, f_{x_{t_1},..,x_{t_n}}(\alpha_1, .., \alpha_k) = f_{x_{t_1+\theta},..,x_{t_n+\theta}}(\alpha_1, .., \alpha_k) \tag{A.10}$$

Le processus x_t est dit stationnaire au sens large, si :

$$\forall t \in I, m_x(t) = m_x \tag{A.11}$$

$$P(t, \tau) = P(t - \tau) \tag{A.12}$$

où $m_x(t)$ est la moyenne définie par $m_x(t) = \mathbb{E}[x_t]$ et $P(t, \tau) = [(x_t - m_x(t))(x_\tau - m_x(\tau))$ est la matrice d'autocovariance. Un processus stationnaire est de moyenne nulle et son autocovariance ne dépend que de la distance entre deux points dans le temps. Un processus stochastique homoscedastique quant à lui, est par définition un processus dont la variance ne varie pas au cours du temps.

Ces notions sont utilisées au cours de différents travaux présentés dans les sections suivantes, en particulier, la notion d'homoscedasticité est utilisée pour définir un profil de rétrodiffusion d'un instrument lidar dans le chapitre 3.

Ces différentes notions ont permis d'introduire dans ce travail des processus tels que les processus Auto-Regressifs d'ordre p, notés $AR(p)$ et définis par :

$$x(t) = \sum_{i=1}^{p} \alpha_p x(t - p) + \epsilon_t \tag{A.13}$$

Ou ϵ_t est un bruit blanc.

3 Espérance et variance conditionnelles

La densité de probabilité conditionnelle est donnée par :

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} \tag{A.14}$$

La dernière égalité est appelée formule de *Bayes*. De cette définition, il découle les premiers moments conditionnels, que sont l'espérance et la variance conditionnelles définies par :

$$\mathbb{E}[X|Y] = \int_{\mathbb{R}} x f_{X|Y}(x|y)dx \tag{A.15}$$

$$\mathbb{V}ar(X|Y) = \mathbb{E}[X - \mathbb{E}[(X|Y)]^2|Y] \tag{A.16}$$

3.1 Quelques résultats sur les moments conditionnels

Nous présentons quelques résultats importants qui découlent des notions d'espérance et de variance conditionnelles. Les résultats présentés ici on été nécessaires pour développer les méthodes des chapitres suivants.

Pour toute sous-tribu $\mathcal{A} \in \mathcal{F}$, l'espérance conditionnelle,$\mathbb{E}[\cdot|\mathcal{A}]$, suit les propriétés suivantes :

i) si X est intégrable, $\mathbb{E}[X|\mathcal{A}]$ l'est aussi. De même, si $X \leq 0$, alors $\mathbb{E}[X|\mathcal{A}] \leq 0$
ii) $\mathbb{E}[\cdot|\mathcal{A}]$ est linéaire, c'est à dire que $\forall(a, b) \in \mathbb{R}$, X et Y deux variables aléatoires réelles : $\mathbb{E}[aX + bY|\mathcal{A}] = a\mathbb{E}[X|\mathcal{A}] + b\mathbb{E}[Y|\mathcal{A}]$
iii) orthogonalité : $Y - \mathbb{E}[X|Y]$ est orthogonal à X
iv) Soit $\mathcal{T} \subset \mathcal{A}$, alors $\mathbb{E}[X|\mathcal{A}] = \mathbb{E}[\mathbb{E}[X|\mathcal{A}]|\mathcal{T}]$
v) Théorème de l'espérance itérée, : $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$

La variance conditionnelle, quant à elle, suit les propriétés suivantes :
vi) $\mathbb{V}ar[X|Y] = \mathbb{E}[(X - \mathbb{X}|\mathbb{Y})^2|Y]$
vii) $\mathbb{V}ar[X|Y] = \mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2$
iix) Si X et Y sont indépendantes, $\mathbb{V}ar[X|Y] = \mathbb{V}ar[X]$
ix) Relation sur la variance conditionnelle : $\mathbb{V}ar(X) = \mathbb{E}[\mathbb{V}ar(X|Y)] + \mathbb{V}ar(\mathbb{E}[X|Y])$

Dans cette thèse, ces relations sont toutes utilisées dans les développements des algorithmes utilisant le filtre de Kalman. Il est possible d'en trouver des démonstrations dans des ouvrages de tels que [Dalang and Conus, 2008] ou [Evensen, 2009].

Annexe B

Résolution du filtrage et du lissage de Kalman linéaire

Nous présentons dans ce chapitre les équations du filtrage et du lissage de Kalman dans un cadre linéaire. Si la résolution du filtrage est communément détaillée dans la littérature scientifique, celle du lissage l'est généralement moins. Nous la donnons ici afin de plus facilement appréhender les Chapitres 4 et 5 de la thèse. La résolution du lissage se basant sur celle du filtrage, nous présentons tout d'abord cette dernière.

Nous nous baserons sur l'équation d'espace-état suivante :

$$x_t = \phi x_{t-1} + e_t^*, \tag{B.1}$$

$$y_t = h x_t + e_t, \tag{B.2}$$

où la première équation représente l'équation d'état, la seconde représentant l'équation d'observation, e_t^* et e_t suivent respectivement des lois Gaussiennes $\mathcal{N}(0, \Sigma_{e_t^*})$ et $\mathcal{N}(0, \Sigma_{e_t})$ orthogonaux. L'objectif du filtre de Kalman est l'estimation des quantités $\mathbb{E}[x_t|Y_t]$ et $\mathbb{C}ov[x_t|Y_t]$, $\forall t \in [1, T]$, avec $Y_t = (y_1, \ldots, y_n)$. Qui représente la meilleur estimation au sens quadratique, de l'estimation de l'état du système à un instant t, au regard des observations disponibles à ce même instant. Le lissage de Kalman quant à lui, a pour objectif d'estimer la meilleure estimation, toujours au sens quadratique, de l'état du système à un instant t, au regard de l'ensemble des observations disponibles, à savoir les quantités : $\mathbb{E}[x_t|Y_n]$ et $\mathbb{C}ov[x_t|Y_n]$, $\forall t \in [1, T]$. Ces estimations sont calculées de manière récursive à partir de l'étape de prédiction, qui correspond aux calculs de $\mathbb{E}[x_t|Y_{t-1}]$ et $\mathbb{C}ov[x_t|Y_{t-1}]$. On définit $\hat{x}_{t,t'} = \mathbb{E}[x_t|Y_{t'}]$ et $\hat{\Sigma}_{t,t'} = \mathbb{C}ov[x_t, x_{t'}]$.

1 Résolution de la prédiction de Kalman

$\forall t \in [1, T]$, par la linéarité de $\mathbb{E}[.|Y]$, la bilinéarité de $\mathbb{C}ov[.|Y]$ et l'orthogonalité des bruits e_t^* et e_t. On estime l'étape de prédiction par :

$$\hat{x}_{t,t-1} = \mathbb{E}[\phi x_{t-1} + e_t^* | Y_{t-1}] \tag{B.3}$$
$$= \phi \hat{x}_{t-1,t-1}, \tag{B.4}$$
$$\Sigma_{t,t-1} = \mathbb{C}ov[\phi x_{t-1} + e_t^* | Y_{t-1}] \tag{B.5}$$
$$= \phi \Sigma_{t-1,t-1} \phi' + \Sigma_{e_{t-1}^*}. \tag{B.6}$$

2 Résolution du filtrage de Kalman

$\forall t \in [1, T]$, on recherche ici à exprimer $\hat{x}_{t,t}$ en fonction de $\hat{x}_{t-1,t-1}$ ainsi que $\Sigma_{t,t}$ en fonction de $\Sigma_{t-1,t-1}$. Cette relation est donnée par :

$$V = y_t - h\hat{x}_{t,t-1}, \tag{B.7}$$
$$S = h\Sigma_{t,t-1}h' + \Sigma_e, \tag{B.8}$$
$$K = \Sigma_{t,t-1}h'S^{-1}, \tag{B.9}$$
$$\hat{x}_{t,t} = \hat{x}_{t,t-1} + KV, \tag{B.10}$$
$$\Sigma_{t,t} = \Sigma_{t,t-1} - KSK'. \tag{B.11}$$

En combinant les deux dernières équations de ce système avec les équations de prédictions précédentes, on obtient la relation recherchée, qui permet d'estimer $\hat{x}_{t,t}$ et $\Sigma_{t,t}$ par récurrence :

$$\hat{x}_{t,t} = \phi \hat{x}_{t-1,t-1} + KV, \tag{B.12}$$
$$\Sigma_{t,t} = \phi \Sigma_{t-1,t-1} \phi' + \Sigma_{e_{t-1}^*} - KSK'. \tag{B.13}$$

3 Résolution du lissage de Kalman

$\forall t \in [1, T]$, on recherche ici à exprimer $\hat{x}_{t,T}$ en fonction de $\hat{x}_{t+1,T}$ ainsi que $\Sigma_{t,T}$ en fonction de $\Sigma_{t+1,T}$, notons qu'il s'agit d'une récurrence d'en le sens *inverse* du temps. La

relation est donnée par :

$$\Sigma_{t+1,t} = \phi \Sigma_{t,t} \phi' + \Sigma_{e_t^*}, \tag{B.14}$$

$$C = \Sigma_{t,t} \phi \Sigma_{t+1,t}^{-1}, \tag{B.15}$$

$$\hat{x}_{t,T} = \hat{x}_{t,t} + C[\hat{x}_{t+1,T} - \phi \hat{x}_{t,t}], \tag{B.16}$$

$$\Sigma_{t,T} = \Sigma_{t,t} + C[\Sigma_{t+1,T} - \Sigma_{t+1,t}]C. \tag{B.17}$$

Annexe C

MEP package : Multivariate Extraction Procedure

Package 'mep'

November 11, 2010

Version 0.4

Date 2008-09-22

Title Multi Extraction of Pulses

Author Julien Gazeaux <julien.gazeaux@latmos.ipsl.fr>

Maintainer Julien Gazeaux <julien.gazeaux@latmos.ipsl.fr>

Depends splines, MASS, fields

Description This package extracts pulse-like events from several series using a Multiprocess Kalman Filter with Fixed Interval Smoother. Extracting Commmon Pulse-Like signals from multivariate Time Series.

License >2

R topics documented:

`mep` *Multi Extraction of Pulses Extracting Commmon Pulse-Like signals from multivariate Time Series*

Description

The program introduce an automatic procedure to estimate the magnitude of strong but short-lived perturbations such as large explosive volcanic eruptions in a series of climate/ proxies time series. Our extraction algorithm handles multivariate time series with a common but unknown forcing. This statistical procedure is based on a multivariate multistate space model and it can provide an accurate estimator of the timing and duration of individual pulse-like events from a set of different time series. It not only allows for a more objective estimation of its associated peak amplitude and the subsequent time evolution of the signal, but at the same time it provides a measure of confidence through the posterior probability for each pulse-like even

Usage

```
mep(X = M, FigFolder="", FNResult="", datacol = c(), timecol = c(), first = c(),
```

Arguments

X	Data matrix containing the data to perform the multi extraction procedure. The columns of the matrix consist of the independent time vectors, and the dependent observed data vectors.
FigFolder	if Plot=TRUE, Character Where to save the output file
FNResult	if Plot=TRUE, character output file name
datacol	index of data column in X
timecol	index of time column in X
first	first index of X from which the calculation will begin. If first=rep(0,n), the data is taken from the start of each of the n time series
last	last index of X from which the calculation will end. If last=rep(0,n), all of the data for the n time series is taken.
lambda	Smoothing parameter for the cubic spline trend for each series. This vector must equal the number of data columns in the matrix X. The lambda value is typically (but not necessarily) in (0,1]
sigmay	A numeric value used in the estimation of the number of pulse-like signals. The default is sigmay=3. Everything beyond sigmae*sigmay is considered a pulse-like signal
Piy	Probability above which pulses will be concidered as real peaks
NumEr	Max number of peaks to be detected
Cat	If TRUE, details of parameter before and after the extraction are displayed during the run
Plot	If TRUE an output file is created to summarized the results

Details

Multi extraction of pulses is a generalization of PulseLikeExtract package. The observations yi(t) are generated by two equations, the observation equation and systems equations.

Model: ..

Observation Equation: yi(t) = xi(t) + fi(t) + ei(t) , where xi(t) describes the pulsatile series, fi(t) describes the trends, and ei(t) N(mean=0,sd=sigmaei) represents white noise.

Systems Equation: xi(t) = alphai*xi(t-1) + v(t) , where v(t) N(mean=mu,sd=sigmae) if I(t)=1 v(t)=0 if I(t)=0 and Pr(I(t)=1)=Piy.

Value

Describe the value returned If it is a LIST, use

`Param`	Estimated parameters of the model, including the number of pulse-like events, Piy, mu, sigmaw,alpha,sigmae,lambda, and a.
`Result`	Different Time series exctracting from input Matrix consisting of t the independent "time" vector, estimated observed series for each of the data series, estimated trends for each of the series, single estimated pulsatile series x, estimated pulse input v, estimated noise for each of the series, and estimated posterior probability of the pulse-like events p

...

References

Gazeaux et al., Extracting common pulse like signal multivariate time series. submitted to Computational Statistics and Data Analysis.

Examples

```
nbdata=3;Pi=0.03;alpha=0.7;sigmav=1.21;muv=4;sigmae=c(15,20,10);beta=c(20,15,7.5);nn=500;P
################################
#     Storage Matrices  for Simulated Data
tim = vector(length=nn) ;           #colnames(tim) = "t"
y = matrix(nrow=nn,ncol=nbdata);    #colnames(y) = paste("y", c(1:nbdata), sep="")
v = vector(length=nn+1) ;           #colnames(v) = "v"
x = vector(length=nn+1) ;           #colnames(x) = "x"
f = matrix(nrow=nn,ncol=nbdata) ;   #colnames(f) = paste("f", c(1:nbdata), sep="")
e = matrix(nrow=nn,ncol=nbdata) ;   #colnames(e) = paste("e", c(1:nbdata), sep="")
occurence = matrix(nrow=nn,ncol=1); #colnames(occurence)=paste("o",sep="")
TS<-matrix(ncol=(2*nbdata),nrow=nn)
################################
#      Starting x value
x[1] = 0
############################
#Must be complete
f[,1] = 10 + 15*sin(2*pi*((1:nn)-1)/90.)
f[,2] = c(1:nn)*.5
f[,3] = rep(0, nn)
################################
#     Simulated Data
u <- runif(nn)
for ( i in 1:nn){
        tim[i] = i
```

```
        if (u[i]<=Pi){
                occurence[i] = 1
                v[i+1]=rnorm(1,mean=muv,sd=sigmav)
        }else{
                occurence[i] = 0
                v[i+1]=0
        }
               x[i+1] = alpha*x[i] + v[i+1]
}
x=x[-1];
v=v[-1]
for(j in 1:nbdata){
        for ( i in 1:nn){
                e[i,j] = rnorm(1,mean=0,sigmae[j])
                y[i,j] = beta[j]*x[i] + f[i,j] + e[i,j]
        }
}
coln<-rep("",(2*nbdata))
TS[,(2*(1:nbdata)-1)]<-tim ;                    coln[(2*(1:nbdata)-1)]  <-c(paste("t",(1:nb
TS[,(2*(1:nbdata))]<-y[,(1:nbdata)];    coln[(2*(1:nbdata))]<-c(paste("y",(1:nbdata),sep="
TS<-ts(data=TS,names=coln)
RES<-mep(X=TS,datacol=c(2,4,6),timecol=c(1,3,5))
##if plot.mep has been sourced
plot.mep(RES=RES)
```

`plot.mep` *Display figure of the output of mep program*

Description

Split graphs in groups (Time series and estimated Trend),(Pulse-probability and Amplitude),(residuals and qq plot)

Usage

```
plot.mep(Folder = "", RESfile = "", RES = "")
```

Arguments

`Folder`	with RESfile to precise the folder and the name of the output file from mep program
`RESfile`	with Folder
`RES`	alone, is the output list of the mep program

Details

input Folder and RESfile must necessarely be together, whereas RES is a single paramter

Value

no output, except display 1st window display the different time series and there estimated trend together 2nd window display the estimated pulse 3rd window the associated probability 4th the residuals 5th the qq plot of those residuals

Author(s)

Gazeaux Julien

References

Gazeaux et al in Preparation

`SimFormep` *Simulator of Time series for mep program*

Description

Simulator of time series correpsonding to state sapce model use in mep program

Usage

```
SimFormep(nbdata = 3, Pi = 0.03, alpha = 0.7, sigmav = 1.21, muv = 4, sigmae = c(
```

Arguments

`nbdata`	Number of time serie
`Pi`	probability of appearance of a pulse
`alpha`	decreasing paramter of pulse amplitude
`sigmav`	standard deviation of pulse amplitude
`muv`	mean of pulse amlpitude
`sigmae`	standard deviation of the noise of the observations
`beta`	relative amplitude of intensity
`nn`	Number of observation per time series
`Plot`	if TRUE, a pdf-file with result pictures is created in 'Folder'
`Folder`	Folder where is stored the result file

Value

`Result`	List with the Parameters (listParam) used to simulate de data (TS)

Author(s)

Julien Gazeaux